The Dissertation Committee for Kanoelani Takaishi Pilobello Certifies that this is the approved version of the following dissertation:

Glycomics: Integration of Lectin and Gene Expression Microarray Data

Committee:

Eric Anslyn, Supervisor

Lara Mahal, Co-Supervisor

Claus Wilke

Vishwanath Iyer

Orly Alter

Glycomics: Integration of Lectin and Gene Expression Microarray Data

by

Kanoelani Takaishi Pilobello, B.S.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

August 2011

Dedication

To my mother who taught me to be free and
my sisters who inspire me to stay that way.

Acknowledgements

I would like to acknowledge my advisor, Dr. Lara Mahal, for all of her patience in training me. I would like to thank my committee members, Drs. Orly Alter, Vishwanath Iyer, Claus Wilke, and Eric Anslyn for their helpful insight over the years. I have been very lucky to have a number of strong female role models. In addition to the women mentioned above, Francis Ligler and Kim Sapsford and Joan de la Cova greatly influenced my desire to stick with it and to understand its importance beyond personal accomplishment. My undergraduate mentors, Rachel Austin, Matt Cote and Sanford Freedman at Bates College, always encouraged me to keep reaching. I also owe my gratitude to the men and women who dedicated their lives as educators in the Chicago Public School system. Many times, I've reflected upon my life and how it cannot compare to the hardship and sheer volatility that my grandparents endured in postwar Japan and the Philippines. To Amadou Cisse, my labmate, who was a mirror of that struggle and who passed as a martyr. To all of my friends, for the hours of coffee, condolences and continued friendship, with specific gratitude to John Reich, Mitra Rana, Michael Weaver, Emily Walter, Bianca Batista, Lakshmi Krishnamoorthy, Luz Carrillo, Jennifer Rumppe, Ana Schaller de la Cova, Melissa Wong, Maren Jimenez, Kendra Curry, Daphne Gomez-Mena, Parvaneh Abbaspour, Emily Potts, Angel Schatz, Alexis Smith, Deepika Slawek, Cynthia Chiong, Liza Eyster and Luis Caridad. It has taken a particularly large village to raise me. So thank you villagers and thank you graduate school pact.

Glycomics: Integration of Lectin and Gene Expression Microarray Data

Kanoelani Takaishi Pilobello, Ph.D.

The University of Texas at Austin, 2011

Supervisor: Eric Anslyn and Lara Mahal

Glycomics is the systematic study of glycosylation in the context of a whole cell or organism. Glycosylated proteins are estimated to make up 50% of all proteins and cover the outside of the cell. Functional roles in glycosylation have been noted in pathogenesis, metastasis, and embryogenesis. However, the structure of these carbohydrates has been difficult to study due to the chemical nature of carbohydrates. Lectins, carbohydrate binding proteins excluding antibodies and enzymes, can be utilized to study glycosylation in a high throughput manner using a microarray format. Glycans, the carbohydrates attached to a protein or lipid, are not synthesized from a template. They are added co- or post-translationally by a concerted set of enzymes in the secretory pathway. In addition, the glycan structures may be altered by metabolism or trafficking.

Cell type specific glycosylation has long been hypothesized due to observations of bacteria homing to tissues. We use lectin microarray technology to define the glycosylation in a subset of the NCI-60, a set of cell lines from different tissues. Using a customized gene expression microarray, we identify cell type dependent glycosylation genes and observe evidence of cell type

dependent spliceforms for an O-glycosylated mucin. Data from the lectin microarray and a published gene expression data set were integrated using Generalized Singular Value Decomposition (GSVD), a linear matrix decomposition method. We have successfully decomposed the data into 3 cell type dependent meta patterns that segregate by glycosylation family. Correlation projection of the genes and subsequent gene ontology enrichment suggests that genes in different pathways covary with the types of glycosylation. An inverse relationship was revealed for the N- glycosylation pattern between the SVD of the lectins and the GSVD of the genes and lectins together. Whereas, the relationship was correlative for O-glycosylation, which was clearly illustrated in biplots. This work argues that types of glycosylation are regulated by different mechanisms in different cell types.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Glycomics is the study of glycans, complex carbohydrate structures bound to proteins and lipids on the cell surface, at a systems level. An estimated 50% of all proteins are glycosylated[1]. Changes in glycosylation have been observed in tumor cell metastasis, embryogenesis and pathogenesis[1-3]. Well established roles of the glycans include ensuring proper protein folding and stabilizing a ligand for binding[4]. Recently, a number of exciting studies have defined functional roles, beyond supporting roles for proteins, for the glycans themselves[5,6].

Glycomics faces two challenges distinct from those of genomics and proteomics. First, individual monosaccharides, with a few exceptions, are chemically identical isomers composed of the same number of carbon, oxygen and hydrogen atoms. In addition, the manner in which two monosaccharides are linked together to form branched, non-linear glycans can vary in structurally distinct ways. These structural aspects can be biologically important. The avian flu virus prefers sialic acid with an α2,3 linkage to galactose which is found at high levels in the bird gut. Human lungs, on the other hand have an α2,6 linked sialic acid preventing invasion[7]. These qualities historically made glycans difficult to analyze by the available analytical methods. Secondly, glycans are produced without a template, unlike RNA and proteins. The branched oligosaccharides are synthesized by addition of a monosaccharide to a substrate oligosaccharide by transferase enzymes. The biosynthetic pathway for N-glycans in which a precursor is co-translationally attached to a protein at an asparagine (Nitrogen) has been modeled under static conditions as a modular

process through the secretory pathway[8]. The pathway for O-glycans which are post-translationally attached to proteins at a serine or threonine (Oxygen) has not been elucidated, however, dynamic regulation and translocation of transferases involved in O-glycosylation during EGF activation has been demonstrated[9]. In addition, the availability of sugar precursors can also influence glycan structure at the cell surface[10]. Thus, there are many factors that influence the regulation of this understudied class of biopolymer.

GLYCAN DIVERSITY

With a few exceptions, the monosaccharides that glycan are composed of are chemical isomers of each other. These monosaccharides are linked together though the anomeric carbon, adjacent to the hemiacetal oxygen at position 1, either equatorially or axially by a glycosidic bond (α or ß conformation, respectively). In addition, the monosaccharides could be linked to one or multiple of the 3, 4, or 6 carbon positions to create branched structures. In the case of terminal sialic acid, linkage can occur at position 8.

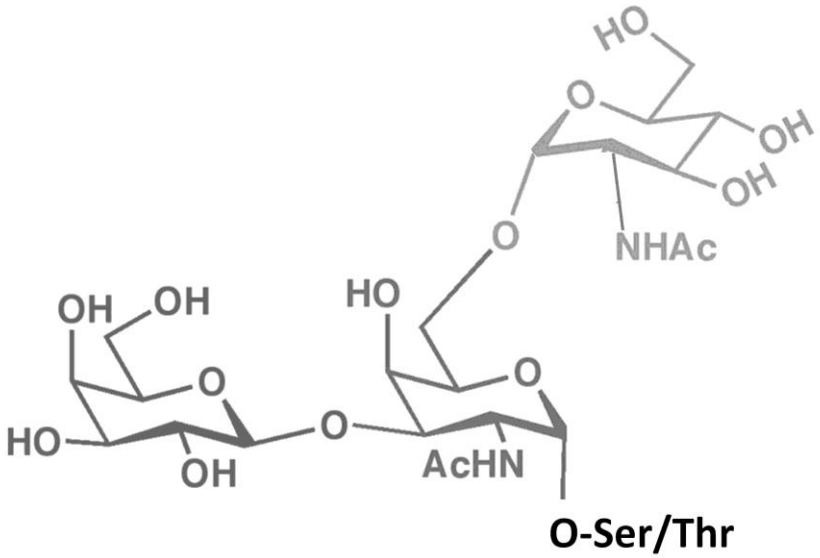


Figure 1.1: Illustration of an O-glycan with 2 conformations. Sialic acid which is modified by an N-acetyl group is also shown.

Additional compounds, such as phosphate and sulfate, may also modify the glycans and affect their function. Laine performed a simple calculation of the number of possible permutation for a hexamer of each type of biopolymer. The results were 4096 hexanucleotide permutations, 64 million peptide permutation, and 1.44×10^{12} hexasaccharides taking the linkage and branching capabilities of carbohydrates into account[11]. Whether or not all of these structures occur in nature is unknown due the technical difficulty of analyzing glycans, however, some estimations have been made.

Based on the number of carbohydrate enzymes, which are considered highly specific, an estimate was made that no more than 500 distinct glycan

structures existed in nature. However, the number of known glycans recently exceeded that number suggesting either unknown carbohydrate enzymes… Cummings estimated the number of glycan determinants, or informative units of carbohydrates, as 7000[11]. This estimate was not only based carbohydrate enzymes, but on the carbohydrate binding specificity of any carbohydrate binding protein including lectins. The impetus to predict glycans from the carbohydrate transferases comes from the fact that there are so many distinct and highly specific transferases. In humans, there are over 300 carbohydrate biosynthetic enzymes[12]. The enzymes in N-glycosylation are largely conserved in mammals (even down to yeast) in terms of sequence. However, glycan composition is lineage specific. For example, yeast glycans are mostly composed of mannose whereas humans have complex glycan structures (Fig. 1.2).[13]

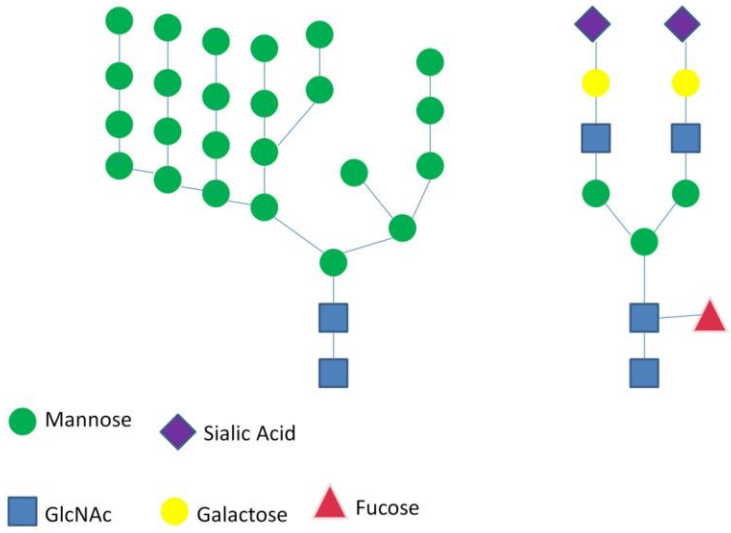


Figure 1.2: A typical yeast high mannose glycan is shown on the left which can be compared to the mammalian complex glycan on the right.

The conservation of glycan synthesis enzymes by sequence does not adequately reflect their diversity. A recent examination showed that some of the enzymes have lineage specific paralogs[14]. In yeast, as expected, the mannose transferase enzyme is expanded compared to other species. In mammals, this is the case for the polypeptide GalNAc transferases involved in O-glycosylation.

Recently genome wide association studies have unveiled significant consequences for dysregulation of glycosyltransferases. Two mutant polysialic transferases were discovered in a patient with schizophrenia. Polysialic acid has been well studied in embryonic neuronal cell migration. These chains are known prevent intracellular interactions during developmental stages[6]. During embryonic neuronal cell migration, the polysialic acid chains on the neuronal cell adhesion molecule (NCAM) are significantly longer than in adult NCAM. The two mutant polysialic transferases discovered affected the length of the polysialic acid chain *in vitro*. The length of the polysialic acid affected the concentration of neurotrophins bound which could have an effect on the pathophysiology of schizophrenia. Clearly, the wealth of information from genomic wide association data which has only become available with cheaper sequencing technologies will play a large role in understanding glycosylation. Glycomics, however, is still in its infancy as the tools to evaluate these structures are currently under development.

EXISITING TECHNOLOGIES FOR GLYCOMICS

Several technologies have been developed to deal with this problem, including tandem mass spectrometry, frontal affinity chromatography, glycan and lectin microarrays[7,15-18]. Although automated glycan annotation programs have been developed, mass spectrometry is still time consuming due to the sample preparation necessary to distinguish between conformations[17]. Glycan arrays are printed with oligosaccharides and have been used to profile the specificity of lectins[19]. A lectin microarray approach has been developed in our lab and in others to provide a method for rapidly detecting overall changes in glycosylation using carbohydrate binding proteins[16]. Many of the plant lectins are commercially available and range in specificity. Some lectins are highly specific whereas others have multiple affinities[20]. This provides an information dense pattern to infer cell surface glycosylation.

Several groups are developing approaches to systems level glycomics[21-25]. Thus far, most studies have focused on validating a single protein or glycan modification. In the most comprehensive study, qRT-PCR was used to quantify levels of approximately 700 glycosylation related genes and N-glycans in 4 different mouse tissues[23]. In most cases, the transcription levels of the genes corresponded well with expressed levels of the glycan structural element they targeted. For certain pathways, the gene transcript levels were shared across tissues. In contrast, for terminal glycosylation genes, transcript levels varied across tissues. Two examples where the transcript levels were not reflected in the glycan structure were oligomannose in the liver and sialic acid levels in the kidney. The authors propose that enlarged Golgi and Endoplasmic Reticulum

organelles that lead to an excess of untrimmed N-glycan precursors are the cause of the raised oligomannose levels. The discrepancy in sialic acid levels remains unexplained. In addition, the authors were restricted to N-linked glycosylation due to the difficulty in sample preparation of O-linked glycans and lipids for mass spectrometry, their method of glycan structural analysis. Despite the limitations, the study provided promising evidence that gene transcript levels will largely explain N-glycosylation levels.

With these considerations in mind, we decided to pursue a glycomics approach using a whole genome expression microarray and the lectin microarray technology developed in our lab to examine the regulation of glycosylation in different cell types. To do this, we chose the NCI-60 cell lines, which are a set of 59 immortalized cancer cell lines distributed by the National Cancer Institute. A wealth of data is available for this cell line set, including mRNA data, miRNA data, DNA fingerprinting, proteomics, copy number variance and chemoselectivity. The cell lines represent 9 different tissues, with multiple cell lines in each type (Table 1.1). We hypothesized that glycosylation is cell type dependent and that we could infer aspects of glycan regulation through the integration of gene expression data. We chose a matrix decomposition method that does not require intensity thresholding to integrate the data due to concerns regarding the low copy number glycogenes. Further considerations and results will be discussed herein.

Cell Type	Cell Lines
Leukemia	CCRF-CEM, HL60, K-562, MOLT-4, RPMI-8226, SR
Lung	A549, EKVX, HOP-62, HOP-92, NCI-H226, NCI-H23, NCI-H322M, NCI-H460, NCI-H522
Colon	COLO-205, HCC-2998, HCT-116, HCT-15, HT29, KM12, SW-620
CNS	SF-268, SF-295, SF-539, SNB-19, SNB-75, U251
Melanoma	LOX-IMVI, MALME-3M, M14, MDA-MB-435, SK-MEL-2, SK-MEL-28, SK-MEL-5, UACC-257, UACC-62
Ovarian	OVCAR-3, OVCAR-4, OVCAR-5, OVCAR-8, NCI/ADR-RES, SK-OV-3
Renal	786-0, A498, SN12C, TK-10, UO-31, ACHN, CAKI-1, RXF-393
Prostate	PC-3, DU-145
Breast	MCF-7, MDA-MB-231, MDA-MB0468, HS578T, BT-549, T-47D

Table 1.1 : List of NCI-60 cell lines

Chapter 2: Lectin Microarray

Technologies that allow for the high-throughput assessment of glycosylation are critical to a systems-based examination of the glycome. Accordingly, our laboratory developed lectin microarray technology as a new method for glycomic analysis. The lectin microarray has been used to characterize the glycosylation of whole cells, bacteria, virions, and tissues. This chapter will provide a brief history of the lectin microarray with a primary focus on optimizing methods for piezoelectronic printing and developing protocols to increase versatility and speed of analysis.

LECTIN MICROARRAY TECHNOLOGY

Lectins are proteins, exclusive of antibodies and enzymes, that bind to carbohydrates. They are comprised of several families and have affinities in the micromolar to mM range[20]. Although this is low by comparison to antibodies (μM – nM), biologically relevant lectin and carbohydrate interactions are multivalent, resulting in higher apparent affinities. For example, multiple viral hemagluttinins bind to multiple carbohydrates on the surface of a host target cell during viral invasion[26]. With this strength in numbers in mind, it is easier to imagine how small changes in the total, localized carbohydrate landscape or subtle changes in structure via linkage and conformation between two monosaccharides, which themselves are chemical isomers, are functional. One of

the advantages of using lectins in an array format is that the lectins are concentrated in the array spot creating an artificial multivalency.

The lectins printed on the microarray are predominantly commercially available plant lectins, which bind mammalian carbohydrate structures for reasons currently unknown[27]. One theory is that the plant lectins which are highly concentrated in seeds have co-evolved with the mammalian gut carbohydrates to protect the seeds by toxicity. Examples of toxic plant lectins include the Ricin toxin which comes from the Castor bean and the lectin from the elderberry plant. The lectin specificities printed on the microarray can range from highly specific (HPA – terminal αGalNAc) to a range of specificities (WGA – βGlcNAc, sialic acid, GalNAc)[28]. The Helix pomatia (HPA) lectin from edible snails is a great example of a lectin from a non-mammalian source that is highly specific to mammalian lectins. In fact, the HPA lectin has been used in cancer prognosis[29,30]. The binding specificities for some of the lectins on our microarray are broad and there is some redundancy, but in most cases, lectins with similar specificities cluster together. Even with these limitations, we have been successful in forming biological hypotheses and validating them[31]. This is a reflection of how much is left to be discovered in the field. Many of these research avenues were previously impossible due to the technology gap which the lectin microarray has helped to fill.

Lectin microarrays were conceptually inspired by gene microarrays. The power of the gene microarray relies on printing a single gene to a specific location or address allowing for high throughput screens[32]. Their ease and successful use in generating biological hypotheses encouraged the development of antibody, protein, and cell arrays. Lectin microarrays make use of the

infrastructure of supporting technologies related to gene microarrays, such as slide scanners, labeling dyes and image processing. Although gene microarrays and their applications have changed since their invention, much of the out-of-the box analysis involves the assumption of a nearly normal distribution and the ability to synthesize random probes as controls[33,34]. These are two factors that have influenced my treatment of lectin array data. At the time of lectin array invention, we could not say *a priori* that the lectin microarray distribution would be normal due to the number of lectins on the array, multivalent aspects of lectins and the unknown nature of biologically presented carbohydrates.

In our proof of concept work, we printed a panel of nine lectins using a manual arrayer on epoxide slides. We hybridized Cy labeled glycoproteins and inhibited these using 100mM monosaccharide to show that the binding event was glycosylation related[16]. We then expanded our lectin panel and began printing using the ArrayIt SpotBot, a microarray plotter that contacts the slide to deposit the lectin[35]. Spot quality, in terms of the shape distribution of the lectin has a serious impact on the data quality. In early microarray experiments, the "coffee ring" or donut effect was often observed[36]. This refers to spots that have a higher fluorescence along the circumference of the spot. The "coffee rings" approach a bimodal distribution which results in underestimating the true mean fluorescence, high error between the medians of replicate spots, and it is difficult to transform into a normal distribution[37]. While that is an extreme example which can be experimentally remedied with the right print conditions, it illustrates the importance of the signal distribution within the spot. This is particularly important with the lectins which have heterogeneous physical properties which result in lectin dependent signal distributions. This could affect

the error distribution across the set of lectins in an array which has wider implications for testable comparisons between them.

Dual color, ratiometric microarray experiments were an established method for accommodating spot morphology issues[38]. In dual color experiments, a sample is labeled with one dye and the reference is labeled with the opposite dye and the two are then hybridized simultaneously (Fig. 2.1).

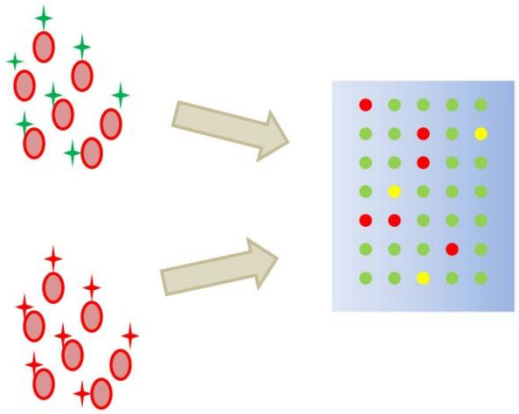


Figure 2.1: Dual color experimental schematics. A sample and a reference are labeled with different dyes and hybridized simultaneously to the same array.

If the experiment is a dye swap, the dyes are swapped in a paired array, so the sample has the dye that the reference had in the first array and vice versa[39]. The dye swap experiment is one of the simplest, least biased methods to account for the difference in dye fluorescence , which is why we chose it. After segmenting the spots with a mask in Genepix, the average of the log2 ratios of each dye swapped pair is clustered and used to compare carbohydrate signatures[40]. In the

dual color method, the carbohydrates of the two samples are in competition with each other. This allows more subtle changes or increases in carbohydrates to be semi-quantified, because of the nature of lectin specificity, in which multiple affinities can be observed. This was the case for wheat germ agluttinin (WGA) lectin binding of a dual color experiment hybridizing both the Lec8 CHO cell line mutant against the Pro-5 parent. When either cell line was hybridized alone in a single color experiment, there was no observable difference between WGA signals. However, with dual color hybridization, we observed the expected decrease in WGA binding for Lec8 which was validated by lectin histology.

M VS. A PLOTS

In addition to the use of dual color experiments to limit spot morphology differences, dye swap experiments allow an experimenter the ability to deal with dye bias. Dye bias occurs because the Cy dyes have different strengths and labeling efficiencies[39,41]. In other words, when scanned at the same gain using photo multiplier tubes, the Cy5 yields lower signal than Cy3. In a gene microarray experiment, it would be easier to adjust for dye differences without a swap because it is easier to define an expectation based on housekeeping genes or other controls. It is possible to adjust the laser powers so that the ratiometric values are 1 for an array where a reference is labeled in each dye and hybridized against itself. However, there can be some slight differences in dye incorporation across a set of samples, so including a dye swap adds some certainty to the observed differences. Dye swap experiments make two assumptions: 1) the

incorporation of dye is the same for all particulates in the sample and 2) there are no other variables that contribute to the dye bias.[40] We accept the first assumption because the proteins and lipids are assumed to be randomly distributed between the cell membrane lamellae. The protocol for making these lamellae includes lysing the cells by sonication which should disrupt any microdomains with particular protein enrichment. To address one aspect of the second assumption, data from a sample labeled with both dyes hybridized to the same array (self-self hybridization) is graphed in an M vs. A plot, where $M = \log_2 (R/G)$ and $A = \log 2\sqrt{RG}$.[42] Spot to spot variation is more easily visualized due to a rotation of the data. In addition, M vs. A plots are useful for determining non-linear affects with increasing intensity. The M vs. A plots for self-self hybridization of 3 different cell lines hybridized at highly separated times were all centered around zero in a straight line as they should be for good arrays (Fig. 2.2). There are some slight differences between the plots. The plot of the p19 cell line experiment shows some labeling differences, which you can see from the two distinct populations. This was a very early microarray hybrization, however, so it is possible that my technique was not good. The plots of the hESC and H9 are differently shaped which suggests that the distribution of signals is different, although they both center around 0. This highlights some of my reservation regarding normalization of lectin array data.

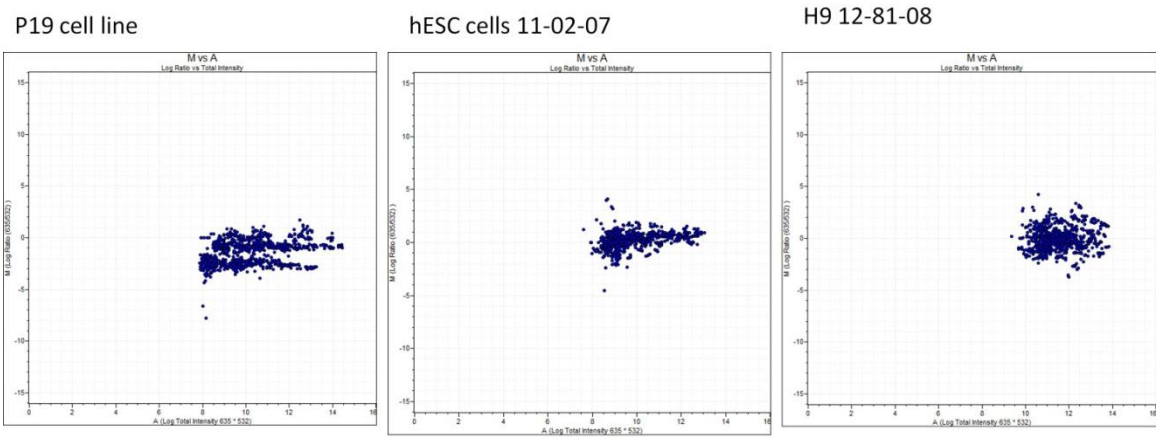


Figure 2.2: M vs A plots of various self versus self dual color experiments hybridized to lectin arrays printed with a contact microarrayer.

PIEZO PRINTING

We recently switched from a contact to a non-contact, piezo plotter (GeSIM NanoPlotter 2.1) which required significant optimization of lectin microarray printing. Microarray plotters dispense probes onto a modified glass slide using a tip or many tips. These tips can either be in contact with the slide surface or dispense by either ink jet or piezoelectronic mechanisms. In both cases, the lectin activity decreases with print time, which should be minimized as much as possible. As there can be print to print variation, it is preferable – though not always possible- to print slides for the same experiment at the same time. Thus, the print time must be optimized. To reduce time using either technology, one must use multiple tips. In the case of contact printing

technology, this limits array formats to the geometry of the printhead. This is not always ideal for multiple array slides and it increases the amount of probe needed. For non-contact arrayers, the faster method is still simultaneous dispensing from both tips, although it is not technically limited to it. Unfortunately, we have found unacceptable differences between lectins printed from different tips and have been trying to increase the print speed in other ways. The considerations for printing with a non-contact printer are quite different from those of a contact printer. One key consideration concerns the voltage and voltage pulse length applied to the tip when dispensing. These parameters can affect the quality of the array alignment and size of the spots. In piezoelectronic printing, the speed of the probes-in-flight is determined by the applied voltage to the tip. The probes-in-flight are small enough that they are sensitive to small fluctuations in air flow requiring that they be dispensed at high enough velocity. Additionally, increasing the voltage to increase the speed and thus the accuracy of spotting, actually influences the size of the spot in a lectin dependent manner. Several of the lectin spots get smaller as the dispense velocity increases, which makes sense given that our slides are coated with a hydrogel (Fig. 2.3). Thus, there is a tradeoff between the accuracy of the spot locations, speed of arraying, and lectin activity.

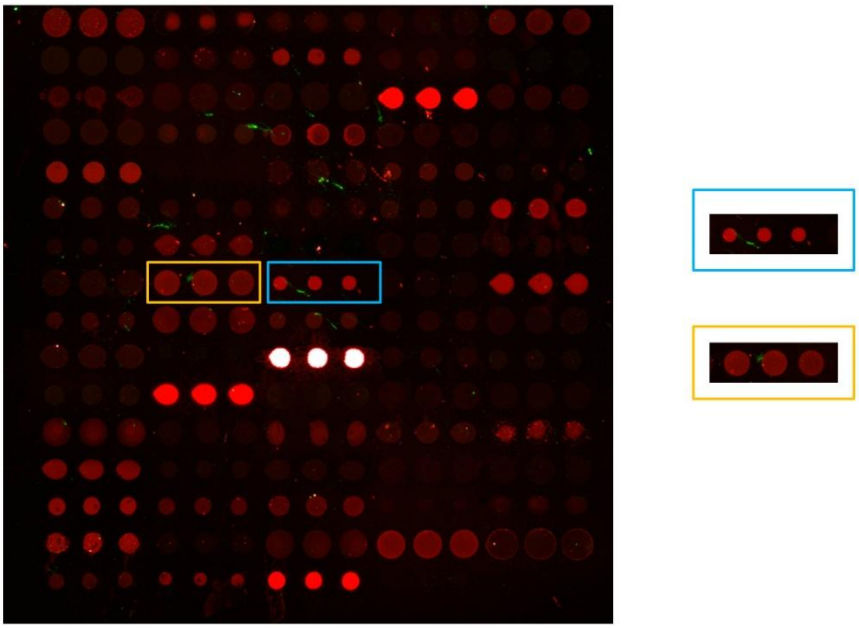


Figure 2.3: A lectin microarray printed with the piezoelectric printing technology at the same voltage. The lectins vary in size across the array. Two different lectins are highlighted in blue and orange.

To compare the effect of voltage and pulse on different lectins and between two different tips, I printed WGA onto the first 12 blocks and DSA on the last 12 blocks (Fig. 2.4a). For two spot rows of 17 columns, 34 spots total, the voltage was varied in increments of 10V starting at 70V, which is the recommended minimum voltage for that size of tip (NanotipA). The pulse was either 50ms or 60ms. Both pins were mounted and dispensed at the same time making every other block row from a different tip. I hybridized one Cy5 labeled glycoprotein to each array, either ovalbumin or asialofetuin.

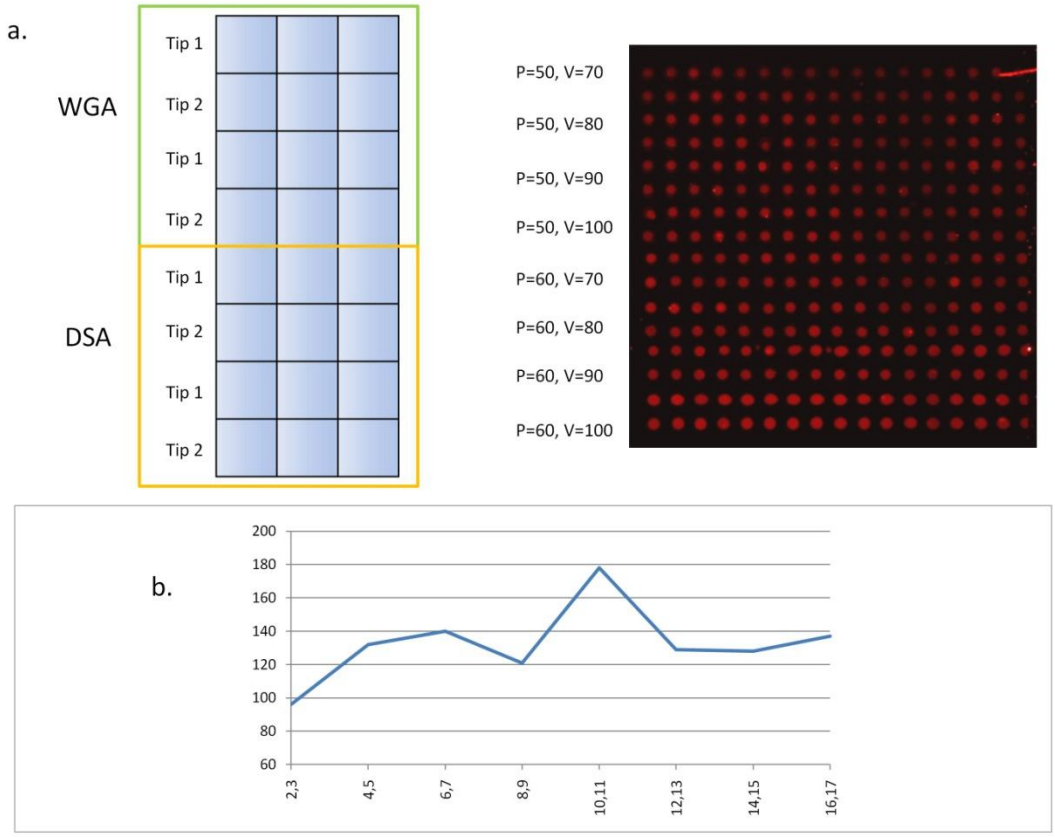


Figure 2.4: Test of print variation by tip. A) schematic of slide layout and an array hybridized with a glycosylated protein. B) Number of spots with high mean-median correlation per set of pulse, voltage conditions.

I used the mean-median correlation, an indicator of the normality of the distribution of fluorescence values and their normality, as a measure of the spot quality[36]. In Excel, the lesser (either the median or the mean) was divided by the greater of the two values. If the quotient was greater than 0.9, the spot was labeled "good," otherwise it was labeled "bad." The number of "good" spots were counted for each set of conditions and compared. The highest number of good spots was 178 for 70V and 60ms pulse and the lowest was 96 spots for 70V

and 50ms pulse. Although the total number of spots for each set was 544, meaning that in the best case on 1/3 of the spots were good, I was using these lectins because they had morphology issues in a previous print. It's possible that the number of good spots is related to the position of the rows since the best set was nearly in the middle at rows 9 and 10 (Fig. 2.4b). This could occur from some affect related to mixing during the hybridization that would favor the middle of the well. However, that trend does not seem to hold, as rows 8 and 9 are actually worse than 6 and 7. Local background would not influence this analysis because I did not use the background subtracted fluorescence values. Comparing the two tips for all pulses and voltages (4352 spots), the new pin printed more good spots than the old pin, 606 and 455 respectively. For the best set (P=60, V=70), a trend of a decrease in percent error going down the slide (Fig. 2.5a). There is a slight decrease inversely associated with fluorescence (Fig. 2.5b), however, the fluorescence intensity and location are not independent because DSA which has higher fluorescence was printed at the bottom of the slide. Given that the M vs A plots for the lectin arrays do not normally show non-linear affects, print location which in this case corresponds to print time appears to have an effect on spot quality. In the practical sense, what matters more is that there is a correlation with printing longer and lower errors. We generally run the printer for overnight washes as recommended, but that does not include dispensing from the tip. In the future, it may help to preprint with water on a different slide before running the print.

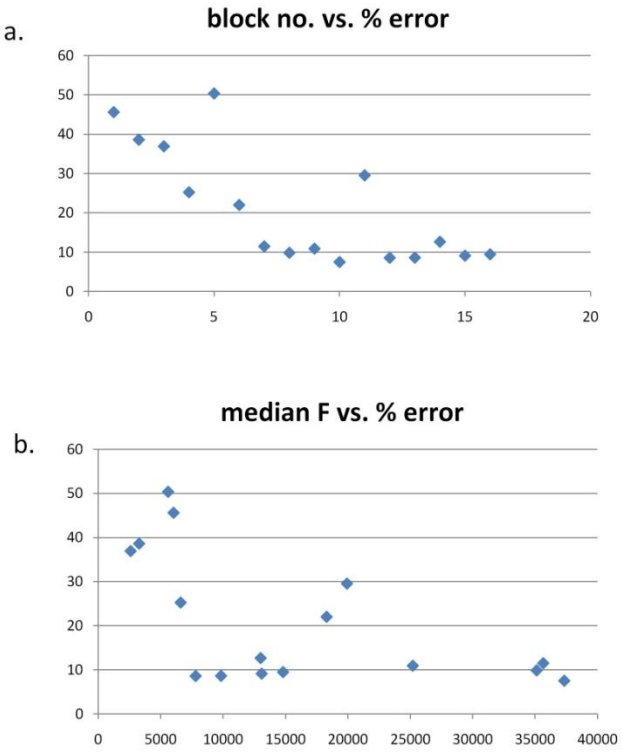


Figure 2.5: Graphs of error against A) block position which corresponds to print time and B) arbitrary fluorescence signal.

THIRD DYE

High throughput is a relative term. While lectin microarrays are higher throughput than many existing technologies for glycan structural estimation, the image and data processing afterward can cause bottlenecks in the workflow. Spot-finding algorithms rely on distinct edges or high spot size and spacing uniformity. In most lectin array experiments, some of the lectins are negative and their edges difficult to distinguish from the background, making them hard to detect. In addition, the flight path of the droplet can be influenced by several variables and the arrays are not always aligned perfectly. Ultimately, a lot of manual alignment is required, which can be time consuming.

A third dye can be added to the print buffer to use for alignment issues.[43] In this case, a positive binding event during hybridization is not needed to detect the spot. The spots from the third dye should also be closer in fluorescence intensity to each other, making it easier to adjust the contrast at which to run the spot-finding algorithm. To investigate the use of a third dye applied to lectin microarrays, I printed one lectin that normally has uniform spots and another lectin that can have irregular spot morphologies (WGA and AOL, respectively). The 5-aminofluorescein diffused into the background (Fig. 2.6). We decided to switch to BSA labeled with Alexafluor-488 instead.

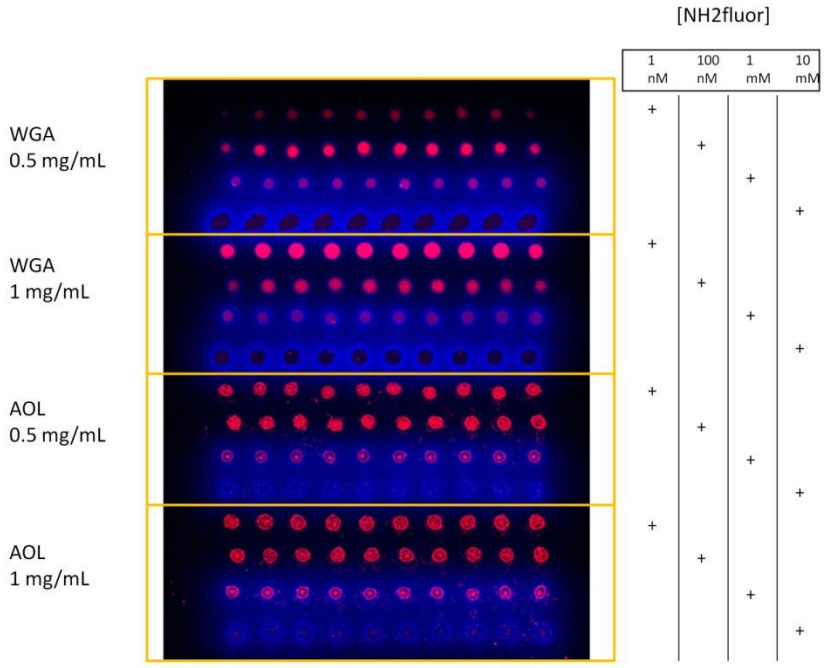


Figure 2.6: WGA and AOL lectins printed with varying concentrations of 5-aminofluorescein dye (blue). The array was hybridized with Cy5 labeled glycoprotein (red).

The Alexafluor-488 labeled bovine serum albumin (BSA) diffused at the same rate as the lectin, which is evident from their overlapping signals. The BSA-A488 spots had strong edges which were easily picked up by the automated spot -finding algorithm, resulting in highly aligned spots. The composite pixel intensity (CPI) is a threshold for determining an edge, where CPI=0 is the most permissive. At a CPI=0, several smears were counted as spots (Fig. 2.7). Whereas at CPI=100, a few spots were missed resulting in 97% alignment. At CPI=500, the properly aligned spots drops to 88%.Based on this data, a CPI between 0 and 100 would minimize the amount of time during manual

alignment although the optimal CPI may need to be determined for each print. Unfortunately, there is some bleedthrough from the Alexafluor-488,. However,this could be fixed with filters with narrower bandwidths or using a different dye.

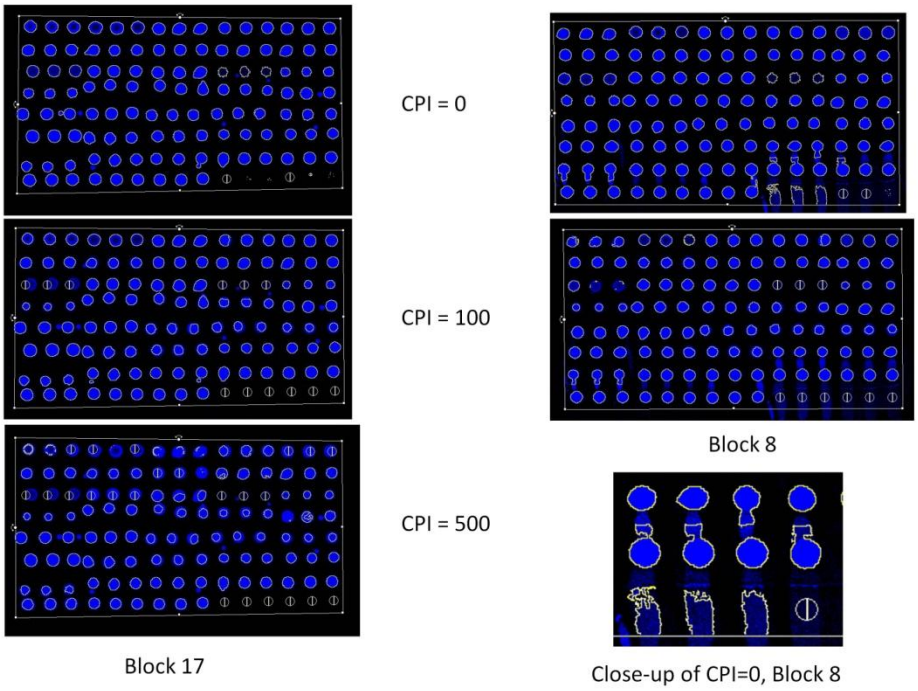


Figure 2.7: Lectin arrays were printed with Alexafluor 488 labeled bovine serum albumin. Automatic alignment can be optimized using the composite pixel intensity in the scanner software.

IMPROVED MEMBRANE PREPARATION

Labeled cell membrane preparations are often hybridized to lectin microarrays. The membranes are isolated by lysing the cells via sonication, ultracentrifugating the solution at 100,000 xg for 1 hour and resuspending the pellet prior to labeling. This process results in heterogenous lamellar blebs. (Fig. 2.8a) Previously, the membranes were isolated after lysing the cells using a probe tip sonicator into the sample that was inserted into the sample and required a volume of at least 1 mL. In addition, the concentration of cells must be at least 1 million cells/mL in order to be above the critical micelle concentration (CMC), the concentration necessary to maintain micellae based on the biophysical properties of the lipids and transmembrane proteins involved.[44] These properties can be extrapolated to lamellae. Since we are working with a complex set of lipids and proteins derived from a whole cell, we had to determine the CMC empirically.

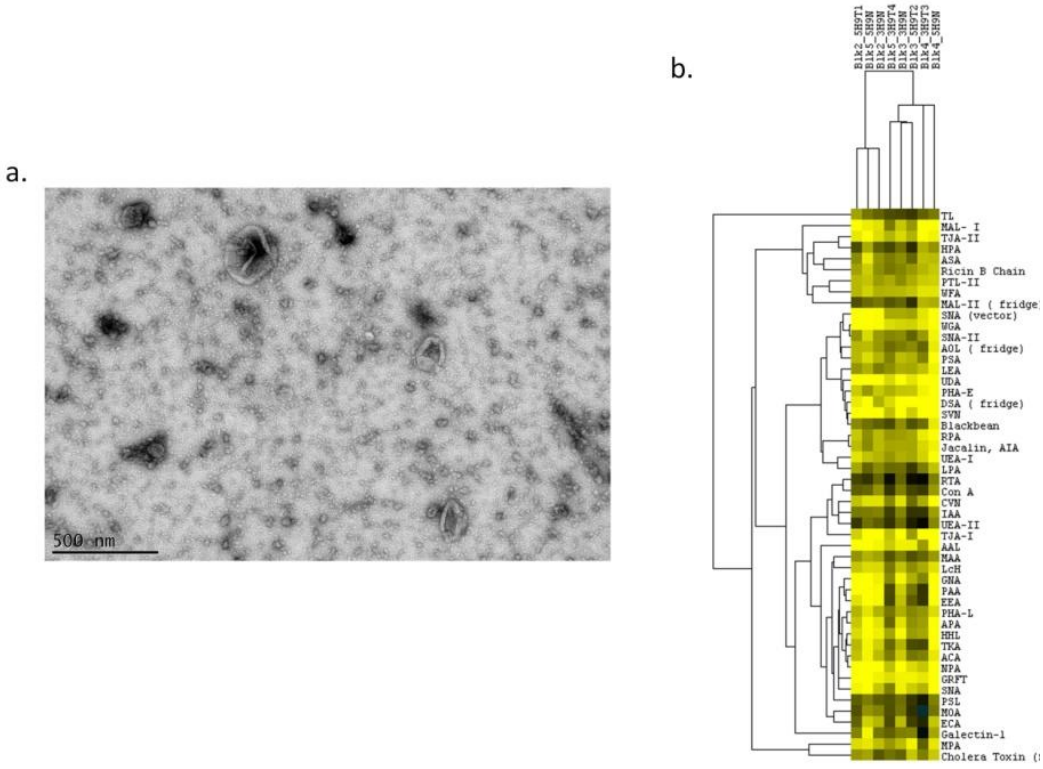


Figure 2.8: A) Electron micrograph of cell membrane preps. Scale is 500nm. B) Cluster of H9 samples prepared using both the Tweeter and probe tip sonicator.

A change in micellar integrity should result in a change in the lectin array pattern because only the proteins are labeled in the micellae, so glycolipid carbohydrate signals would be lost or could possibly compete with the glycoprotein signal. I found that below a concentration of 1 million cells/mL the lectin array pattern was significantly different. Thus, with the probe tip sonicator, we are limited to samples of at least 1 million cells. Certain types of cells, for example, neuronal and primary cell lines, cannot grow to a high cell density. I tested sonication using a device that can sonicate several eppendorf tubes of sample at the same time (Tweeter). I sonicated H9 and Jurkat cells using the Tweeter at a concentration we often use of 10 million cells in 3mL. I varied the pulse time (5, 10 seconds) and the power (50, 100%) and kept the rest time for all samples at 15 seconds. We sonicate with 5 second pulses with 10 second rests as described in the user manual when using the probe tip sonicator. Using cells from the same flask, I also prepared them using the cell membrane preparations using the probe tip sonicator and labeled all the samples at the same time. The probe tip and Tweeter samples clustered at 0.637 for H9 cells. (Fig. 2.8b).

CONCLUSION

Lectin microarrays have enabled our lab to address fundamental questions in glycobiology, and allowed us to engage in fruitful collaborations. Over the years, the lectin technology has improved significantly. The printing technology is now more reliable and effective and additional carbohydrate binders, such as bacterial adhesions although not discussed here, have been

developedfor lectin microarray use.[45,46] Overall, the lectin activity has improved with non-contact, piezoelectronic technology. However, there are some considerations for printing lectins related to the voltage and pulse when using this printing techology, which I hope I have illustrated. In addition, the arrays have expected M vs. A plots and are not subject to severe non-linear effects, so in most cases, depending on the sample, should not require data smoothening. While the technology continues to mature in its application, several issues need to be solved. Among these are optimizing the print time and conditions, resolving data processing bottlenecks, and miniaturizing sample preparation protocols to be useful for low density samples. Also, laser dissected tissues would require a resizing of the microarray.

METHODS

Lectin array printing

Plant lectins (EY Laboratories, Vector Labs, Sigma) are printed in 0.5, 1, and 2 mg/mL concentrations on Nexterion H slides (Schott). We are currently printing with a piezoelectric NanoPlotter 2.1 (GeSIM). The slides are printed at 45% relative humidity on the plotter's cooling deck which keeps the circulating water at 10°C (Fig. 2.9). Ideally, we print the slides in 24 hours, which is approximately the time it takes to print 14 slides with 90 lectins, or less. Prior to printing, the NanoPlotter is placed on standby mode and the tip is washed every 30 minutes overnight.

One of the main advantages of the NanoPlotter is the flexibility of the print conformations and substrates. It is possible to print onto any target and in fact, grab coordinates from an irregular object in an image and print those coordinates into prototype silicon microwells, which we have done. The NanoPlotter also comes with a series of programs for spotting from multiple tips simultaneously or sequentially. However, for ultimate flexibility in source plate aspiration and spotting, the Multitask program can be used. This program was developed by HTS resources to run spotting programs using XML markup code. I have used this to make program that results in an array where a singular spot of lectin is printed after the last lectin rather than printing 3 spots of the same lectin in succession. We did this to distribute the lectins around the slide to better account for background variation. The XML transfer lists command the NanoPlotter to spot in a line by line fashion with source plate, block, and spot

coordinates. I have written multiple scripts in Python to generate these transfer lists.

Another aspect of the NanoPlotter is the Stroboscope test. The Stroboscope takes 4 images of rapidly dispensing drops. Parameters of the drops are measured for consistency. If the variation in the Y location of the drops is too high, the NanoPlotter will not continue printing that sample. At the end of a program, a correction file is run with the coordinates for the missing spots.

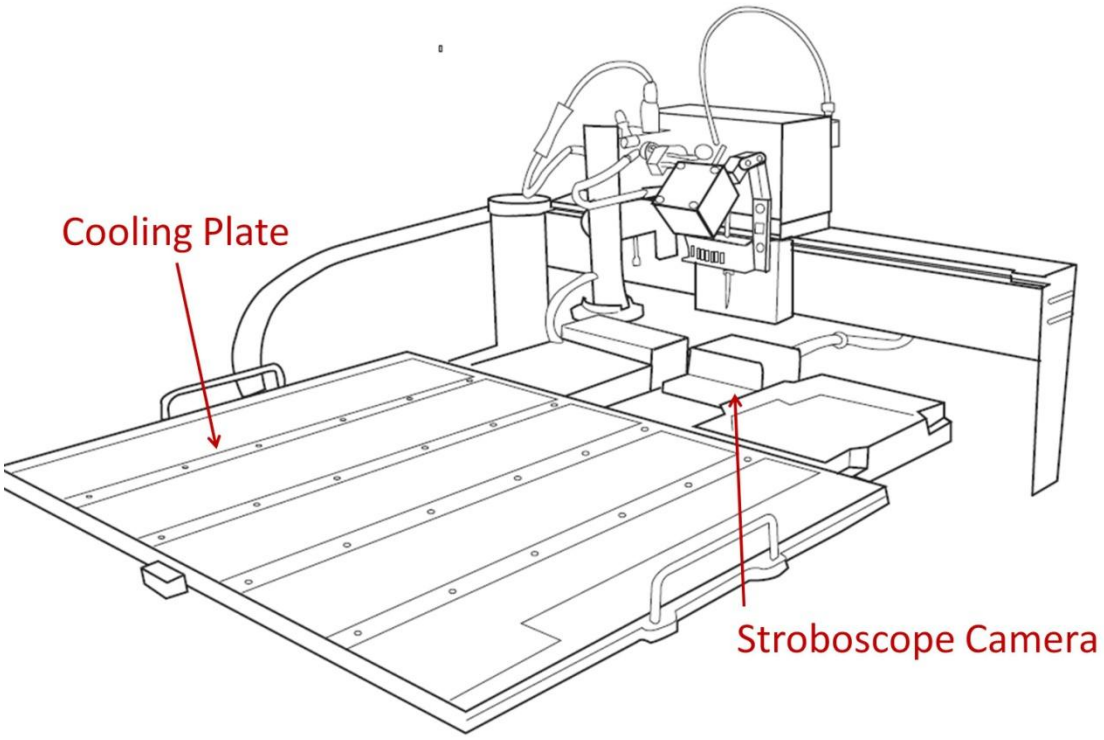


Figure 2.9: The NanoPlotter slide deck and printhead.

Membrane isolation and labeling of mammalian cells

Adherent cells are harvested by scraping to avoid Trypsin which cleaves glycans in a biased way. The cells are then pelleted and resuspended in a phosphate buffered saline at a concentration of at least 1 million cells/mL. The cells are lysed by sonication and the membranes are pelleted from the solution by ultracentrifugation at 100, 000 x g for 1 hour at 4°C. The pellet is resuspended in carbonated buffer (pH = 9.3) for labeling. 6μg per mg of protein of either Cy5 or Cy3 dye (Amersham Biosciences) is added. The dye and the membrane particulates are incubated at room temperature for 30-45 minutes. The reaction can be quenched with Tris buffer. The samples are then dialyzed overnight with a microdialyzer at 4°C. Following dialysis of the free dye, the samples are aliquotted, snap frozen in liquid N_2 and stored at -80°C.

DSA and WGA spot quality assessment print

The arrays were hybridized with 10μg of glycoprotein, either Ovalbumin or Asialofetuin. As per usual, the slide was incubated at room temperature for 2 hours and washed 5 times with PBST (0.05% TWEEN) and once with PBS prior to scanning on our Genepix 4300A. The spots were segmented in Genepix Pro 7.0 using circular alignment with some manual adjustment. The left column of the slide was lower in intensity than expected so I only used the middle and right columns for analysis (16 blocks total). The mean median correlation was used to avoid contribution of background variation in the assessment of the spots. In

addition, the data for the whole slide can be compared without regard to the glycoprotein identity.

General segmentation of data and analysis

The slides are scanned using a GenePix 4300A Scanner (Molecular Devices). The data is segmented from the image file (TIFF) of the scanned slide in Genepix Pro 7.0 using circular alignment. The alignment is manually checked and corrected if necessary. The results are exported in text file and analyzed, for the most part, in Excel. Local background subtracted median fluorescence is tested for outliers using the Grubbs test. Signals are counted as positive if the SNR ≥ 3 in the Genepix report. Dual color experiments are usually dye swap experiments. In dye swap experiments we calculate the average of the Log2 ratios for the pair.[35,40] The averages are hierarchically clustered in Cluster 3.0 using the Pearson centered metric between arrays and average linkage.[47] To cluster the lectins, we use the Euclidean distance.

3rd dye printing

Well defined Alexafluor 488 bovine serum albumin (BSA) was purchased. BSA is a globular protein that is not glycosylated and has been used in protein array print buffers as a passivating agent.[48,49] The Alexafluor BSA was added to the print buffer at a concentration of 100ug/mL.

Tweeter specifics

I put the samples in the positions closest to sonicating source which are the highest power because the resistance increases along the metal holder. The samples were sonicated with 3 pulses at 100% power. I have observed some

variability with using lower power percentages. The length of the pulses and rests (5 seconds and 10 seconds, respectively) are the same as they are for probe tip sonication, which was suggested in the probe tip manual.

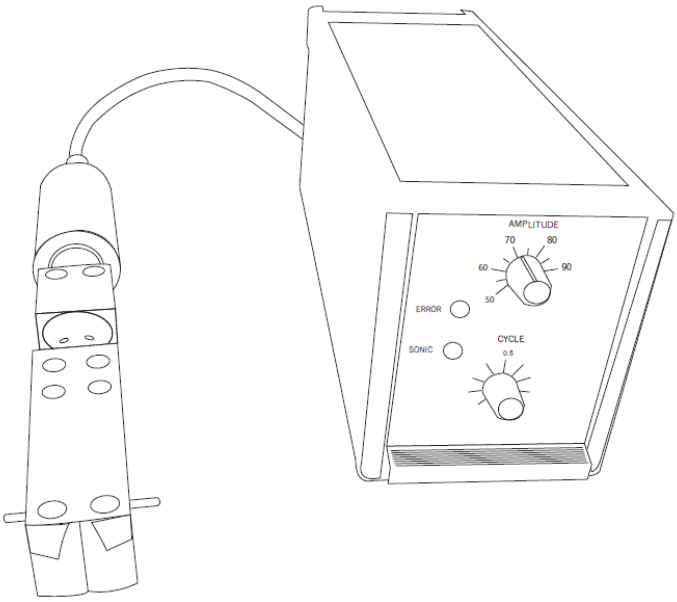


Figure 2.10: Tweeter sonicator. Eppendorf tubes can be placed in the holes of the holder. The tubes closer to the power source experience less resistance and are subject to higher power sonication.

Chapter 3: Whole Genome Expression Array

Gene expression microarrays are highly affordable and robust, and since their invention in the mid 90's, have matured as a technology with clinical utility.[32,50] As the technology transitioned, there were mixed reports about the reproducibility of the experiments of different labs and platforms. Each microarray platform has a different approach to designing the gene expression probes. These probes must match part of the sequence of a gene or predicted gene while allowing for technical aspects of the labeling of isolated RNA, which could result in shortened transcripts. The Affymetrix microarray platform uses many (approx. 20) short probes of 20-25 bases and calls genes based on the average of full matches minus mismatch probes to reduce the contribution of background and cross-hybridization. The Nimblegen platform uses 60mer (bases) probes that are designed to contain 37 overlapping 24mer sequences. In addition, the companies often sell labeling and kits for isolating RNA from samples, which could add variation. The MAQC study was initiated to address previous reports of inconsistencies between labs and platforms and published its first report in 2006[51]. The MAQC studies have shown that for the same sample there is a high level of technical consistency between array platforms and sites despite different labeling protocols. The results suggest that a portion of those discrepancies are real transcript levels and could be biologically relevant. In light of the ENCODE project, it is apparent that there are more and different transcripts than previously predicted.[52]

Gene lists can vary and the coverage for some biological processes is not complete in standard microarray platforms. This is the case for glycosylation. In fact, a custom array targeting glycogenes was curated and designed using the Affymetrix platform by the Consortium for Functional Glycomics.[21] Today, it is possible to affordably buy and even customize expression arrays with probes for 44,000 genes. We chose to study the regulation of glycans in different cell types using the NCI-60 set at the gene transcript level using a customized whole genome expression array. This chapter will discuss my work with available gene expression sets for the NCI-60 and the design of a gene expression microarray customized with a full set of glycogenes.

PRELIMINARY EVALUATION OF EXISTING DATA

Glycosylation as a cell type marker has long been theorized. One observation related to the establishment of this theory is pathogen tropism, or host selection. In fact, E. coli targets different tissues by using fimbriae that target carbohydrates in a specific tissue.[13] The P and type-1 fimbriae target the urinary tract by binding to Galα1-4Galβ- and Manα1-3(Manα6)Man glycoforms, respectively. The E. coli S fimbriae targets neural GM2 and GM3 gangliosides andthe K99 fimbriae bind to intestinal cell wall gangliosides. Carbohydrate changes have also been observed in embryogenesis, when cells begin to identify and distinguish themselves from other types.[1] In Drosophila, cell identities are established initially by a gradient which is essential to proper polarization and can affect the downstream stages when cells are further differentiated.

Significantly, a gradient of a sialic acid analog has been observed in Drosophila larvae.[24] Thus, there are many provocative observations that suggest a cell type dependent sugar code, however there has not been a systematic study of it.

Although many of the details of carbohydrate regulation have yet to be fully understood, there is a fair amount of evidence for regulation of the glycosyltransferases at the transcript level for N-glycosylation.[8,23,53] We decided to look at the contribution of glycogenes to cell type from published gene expression array data. Ross et al. hybridized the NCI-60 to an array with approximately 8000 gene probes.[54] One result of the study was that 1160 out of 8000 genes clustered by cell type. Within this set of genes, the ovarian, leukemia, colon, CNS and renal cells clustered by cell type. We evaluated this based on the clusters below the Pearson Critical Value of 0.062 (P = 0.05). The Pearson Critical Value is a threshold of significance calculated for a particular number of samples and P-value using a t-test. To see how many of those were glycosylation related genes, I extracted out the glycosylation related genes from the whole set of positive signals and compare their clusters to the published cell type clusters. This resulted in 166 glycosylation related genes out of the approximately 4800 genes with a significant fold change.

The glycogene subset was hierarchically clustered. The Pearson critical value was calculated for the smaller number of genes and used as a threshold of significance for the gene clusters (fig. 3.1).

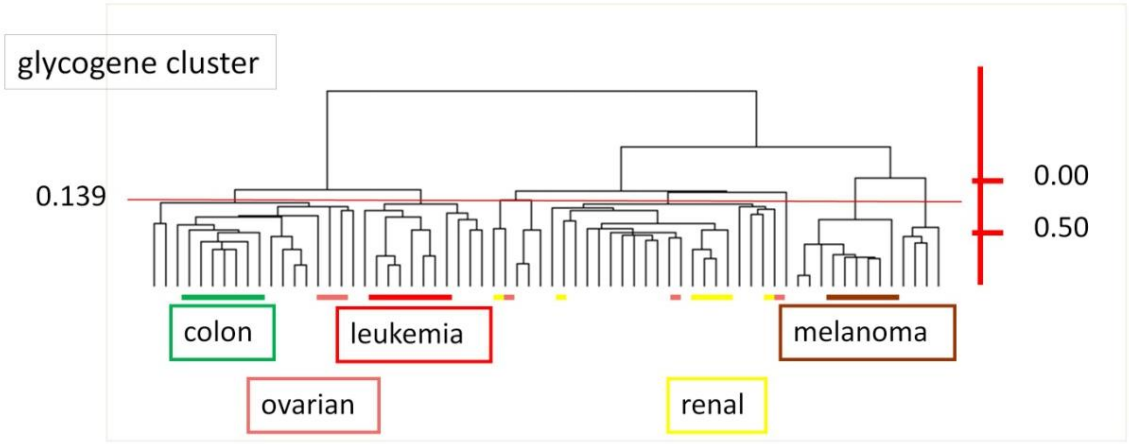


Figure 3.1: Cluster of arrays in Ross data based on glycogenes. The Pearson critical value for 166 genes is displayed.

The colon, leukemia and melanoma cell lines clustered in both the glycogene subset and the original dataset. The renal and ovarian cell lines did not cluster as well in the subset and were in fact, sparsely distributed across the clusters. However, the glycogene subset for these data only included 14% of the 1160 glycosylation related genes which have been curated by the Consortium for Functional Glycomics (CFG). This list includes glycosyltransferases, glycosidases, glycoproteins, lectins, sulfate transferases and other relevant genes.Thus, the underrepresentation of glycogenes might explain the lack of expected correlation between the renal and ovarian cell lines. Another possibility, is that other genes influence the regulation of glycosylation. This has recently been illustrated with regard to a Fucose 1,3 addition which is important in neural development of flies. Using a forward genetic screen in Drosophila for

this addition which results in physically deformed 3rd instar larvae 166 genes relevant to the proper addition of Fucose 1, 3 were identified.[24] These included master regulators, some of which were actually cell cycle dependent genes that have no prior association with glycosylation. Although this paper was published after our experimental design, it provides a promising outlook for the completion of our experiments which will be able to look at multiple glycosylation structures via the lectin microarray.

CUSTOM ARRAY DESIGN

Our objective was to customize an array with all of the glycogenes on the glycogene array (version 4.0) which was curated by the Consortium for Functional Glycomics in addition to the other genes.[21] First, I matched RefSeq IDs from the CFG list to the Nimblegen catalog resulting in 820 matches. A few were no longer active RefSeq IDs due to redundancy. The remaining gene IDs were largely UCSC browser IDs which is a less stringent catalog than RefSeq. RefSeq is a curated database of non-redundant gene sequences submitted to the International Nucleotide Sequence Database Collaboration.[55] It does not include isoforms or splice variants. The Nimblegen whole genome expression array actually includes some redundant genes which may have been identified by different groups. These genes could be on different locations and different parts of the chromosome. For these reasons, Nimblegen does not accept gene name IDs when designing custom probes and only accept chromosome locations. I used the UCSC table browser to retrieve the exon sequences for the UCSC IDs.

Nimblegen selects probes for targets within 1500 base pairs from the 3′ end on the gene. This makes the probe less sensitive to splice variants which generally occur on the 5′ end. Through a complex set of rules related to the uniqueness of the probe and optimal chemical properties, such as melting temperature and unlikelihood of secondary structure, they select the best set of probes for the target gene. Many of the resultant glycogene probes include probes to multiple exons of the same gene.

VALIDATING LOW COPY NUMBER GLYCOGENES

Many glycosylation related genes are low copy number. To insure that that the low copy genes were reverse transcribed, I used both OligodT and random primers in a 1:1 mix.[23] I validated the presence of low copy genes by PCR amplification of the double stranded cDNA. Rft1, Chst14, and St6GalNAc1 were chosen based on their relatively lower expression levels yet presence in all 4 mouse tissues.[23] Rft1 is the gene encoding an endoplasmic reticulum transmembrane protein involved in flipping the lipid linked N-glycosylation precursor into the ER prior to elaboration. It is highly conserved down to yeast and its expression would be expected in all cell types. Chst4 is the gene for a dermatan sulfotransferase and St6GalNAc1 transfers CMP-sialic acid to both N and O linked glycans with an α2,6 linkage. PCR amplification of Rft1, Chst14 and St6GalNAc1 dsCDNA from a total RNA isolation confirmed the presence of Rft1 and Chst14 (Fig. 3.2).

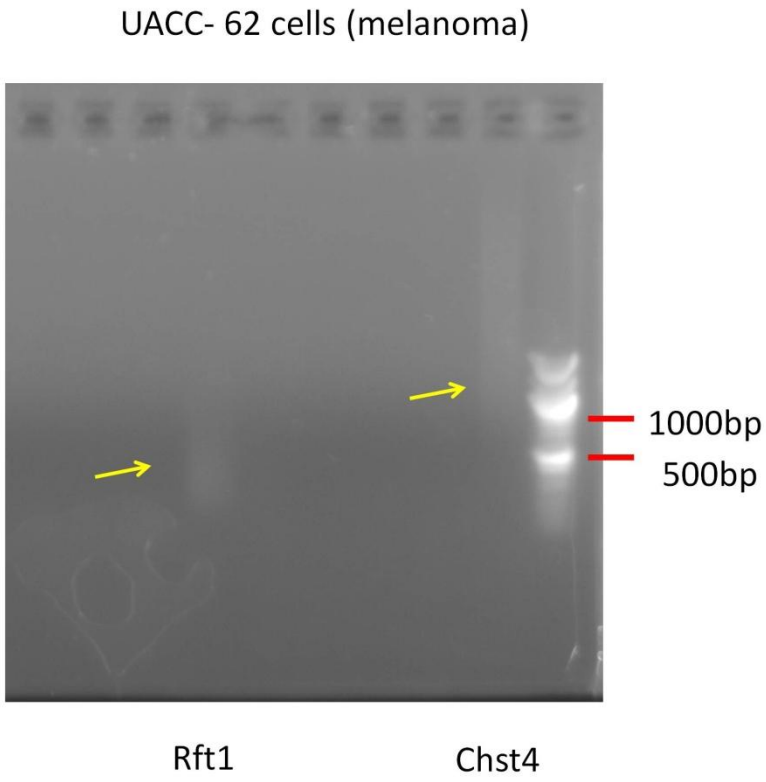


Figure 3.2: Polymerase chain amplification of Rft1 and Chst4 in cDNA from UACC-62.

Although the PCR resulted in a streak, the two streaks were centered at the expected transcript lengths (183 bases for Rft1 and 970 bases for Chst4). Unfortunately, I could not confirm the presence of St6GalNAc1 by PCR although I repeated it several times. The CFG glycogene array list only includes one entry

for St6GalNAc1, so if there are splice variants, they are not currently known. Sialic acid terminates glycan chains, except in the case of oligomeric polysialic acid, and as a family of glycosyltransferases, the sialyl transferases are differentially expressed in different tissues. It's possible that although St6GalNac1 was expressed in the 4 different mouse tissues that this gene is not expressed as widely in humans. Taken together, the results of the Rft1 and Chst4 PCR confirm that low expression transcripts were present in our samples. In fact, the probes targeting all three genes were above background but below the mean expression value in our arrays which were hybridized using the same protocol for cDNA synthesis.

ANALYSIS OF GLYCOGENE PROBES

The data from our gene expression array were examined for cell type dependent signatures and the overall success of the probes. To briefly describe the data collection, matched samples from the NCI-60 were collected and split into cells for the lectin array and cells for the gene expression array. The total mRNA was isolated and labeled cDNA was synthesized. The samples were checked for quality prior to hybridization. The data for each gene array was extracted from the image and the quality of the hybridized array was evaluated using the NimbleScan software.

I selected the glycogenes from the data for the 49 cell lines which passed all quality control measures. The glycogenes consisted of 2994 probe sets, many

of which targeted different exons of the same gene. 64% of the glycogene probes were not overexpressed in any of the arrays (Fig 3.3). This was determined by a stringent cut-off of 2-fold over the mean for each probe. The 2-fold over the mean cut-off was recommended by the MAQC consortium as threshold that reliably identifies biologically important data. This does not exclude the possibility that probes (or genes) below the threshold could be important. However, clustering methods are limited by the amount of computation power required for large data sets. In addition, too many low expressing probes could overwhelm signatures and make the cluster uninformative.

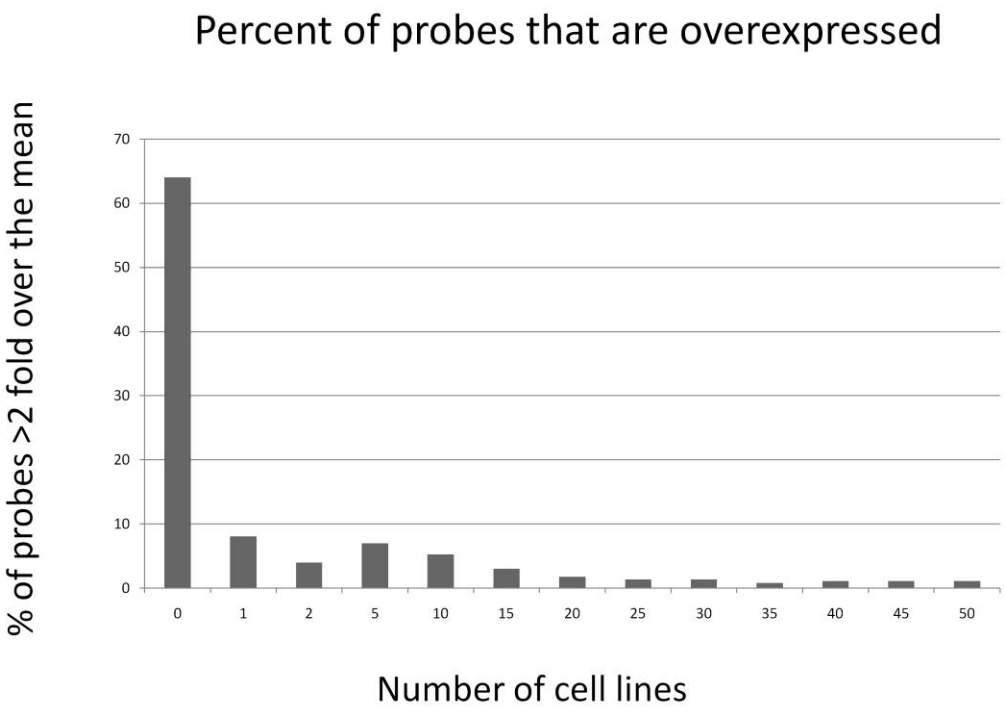


Figure 3.3: Histogram of the number of cell lines that a probe was positive for.

Another 8.0% of the probes were only positive for one sample. 76% of the 241 probes that were positive for on 1 gene, were probes designed to a specific exon of one of the missing glycogenes rather than one of the higher confidence probes provided by the Nimblegen gene catalog. Approximate half of those singular signals were from 2 cell lines, UACC-257 and NCI-H460 (53 and 60 probes, respectively). Although these cell lines passed the uniformity threshold, the difference in their overall expression from the other cell lines is suspicious. They were included analysis for the time being. On the other hand, just 1.1% of the glycogenes were broadly overexpressed in 45-50 samples.

Genes that were shared between 90% of the samples included a number of sulfate transferases and mucins and lacked proteoglycan synthesis genes (Appendix A). Notch 2, Notch4, Jagged 2 and fibroblast growth factor receptors are involved in proliferation, as would be expected for cancer cells. Proteoglycan synthesis is often associated with cell proliferation and angiogenesis as well, but their upregulation was not shared amongst all cell lines. It is possible that their loss was necessary for adaptation to cell culture where cells are grown two dimensionally on a culture dish rather than three dimensionally. The lectin array comparisons of cells grown two-dimensionally in culture plates and three-dimensionally in Matrigel showed no differences in cell surface glycosylation (personal communication, Monika Abramszuk). However, The glycosyltransferases that are upregulated include UGCGL2 and GALNT10, which are a glucosyl ceramide and polypeptide GalNAc transferase, respectively. GALNT10 is involved in mucin synthesis so their coexpression is understandable.

The positive glycogene probes were clustered (Fig 3.4). The whole set was highly correlated (r = 0.59). The leukemia and melanoma cells group together, but because the overall correlation is high, they do not segregate from the rest of the group. The rest of the cell lines are intermixed by type. It is possible that there are still cell type dependent signals that cannot be determined by hierarchically clustering the over 900 positive glycogene probes. I filtered the data by various parameters. Thresholding the data for genes that were positive in five cell lines did not improve the cluster nor did taking the new probes out. Methods other than cluster can be used to evaluate the data, however, I wanted to compare our data to the cluster of the Ross data glycogenes, presented earlier.

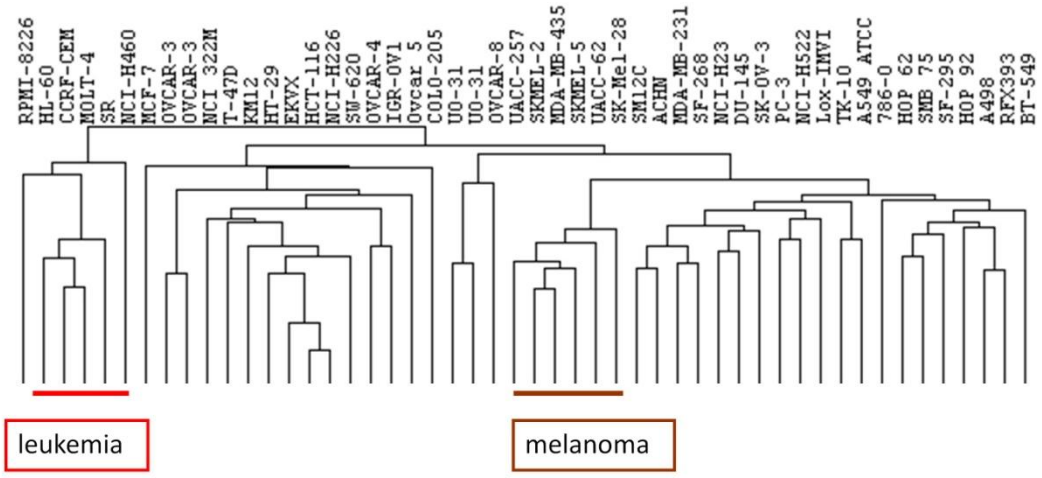


Figure 3.4: Array tree of the clustered glycogenes from the Nimblegen array.

The frequency distribution of the number of cell lines that a gene is positive for shows an overall decline (Fig. 3.3). However, there is a hump for genes that are positive in 3 to 10 cell lines, which is similar to the number of cell lines in one cell type. With the exception of the prostate cell type which are only 2 lines, the rest of the cell types contain 5 to 10 cell lines. I filtered for genes that were positive in 5 to 10 of all 49 cell lines and took a subset of cell lines with distinct gene expression signatures, which was determined from previously published data. The cell lines are the leukemia, colon, melanoma and renal cell types. The genes were ordered by cell type and clustered by gene (Fig. 3.4). Visually, there are some cell type dependent signatures in the heatmap. 72% of the genes in this set are from the probes that we customized.

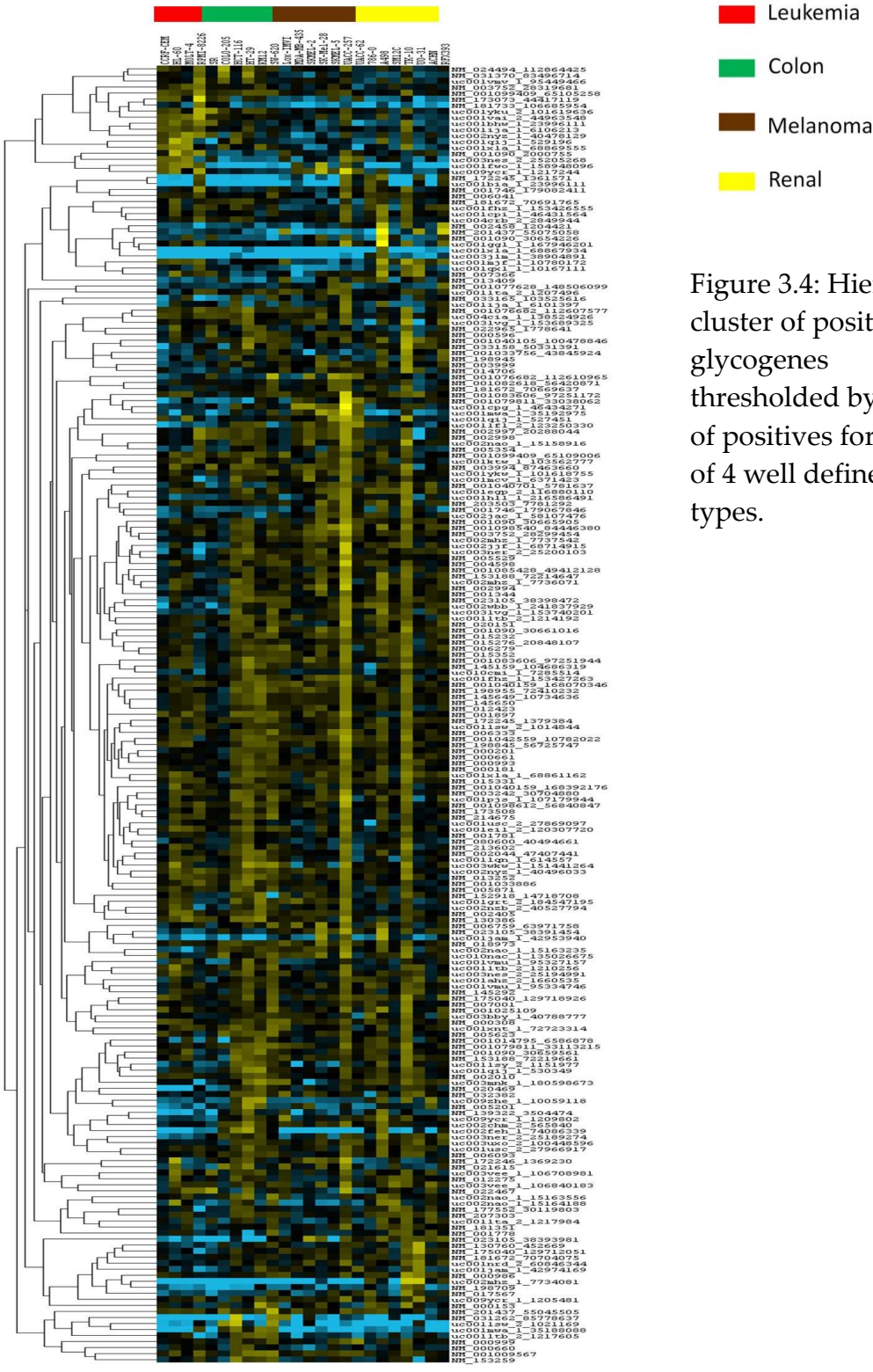


Figure 3.4: Hierarchical cluster of positive glycogenes thresholded by number of positives for a subset of 4 well defined cell types.

To look at these cell type dependent signatures, I counted positives by cell type and filtered the data for genes that were positive in only one or two of the cell types. The genes are segregate to these cell types by class of glycoprotein (Appendix A). The colon and renal cells both have overexpressed C-type lectins (calcium binding) and mucins. Notably, the cell type specific and shared set of C- type lectins and mucins are distinct. The shared mucins include MUC1 and MUC5B. Furthermore, the cell type specific mucins actually segregate between the renal and colon cells. MUC 17 and one of the MUC5B probes are overexpressed in renal cells. Whereas MUC6 and a different MUC5B probe are overexpressed in the colon cells. Interestingly, the MUC5B probes correspond to different probe locations. Two of the probes were custom probes designed from the same gene location, but correspond to two different exons from one of the newly designed UCSC probes. The other MUC5B probe was designed to the more frequently included Refseq probe. Mucin genes have high levels of polymorphism and have been observed in tandem repeats. In fact, several subpopulations of the HT-29 colon cell have been isolated where differences in expression of MUC1 and MUC5 were observed with age (passage number).[56] However, the transcript levels highly correlated to the mucins expressed, suggesting transcriptional regulation. Leukemias generally express simple O-glycans where the first sugar added is GalNAc. The leukemia cells exclusively overexpress sugar transporters. GALE is an epimerase which converts glucose or GlcNAc to galactose or GalNAc. The other two genes do not have well defined functions. SLC35C2 is overexpressed in ovarian cancer and is related to

maintaining oxygen levels. This could be rationalized by the fact that leukemias are blood cells. The last sugar transporter is CMP-NeuAc hydroxylase protein homolog. A loss of the catalytic N terminal of CMP-NeuAc hydroxylase (CMAH) is a human specific mutation resulting in a lack of N-glycolylneuramic acid (Neu5Gc).[57] It was thought that Neu5Gc was not expressed in humans; recently, however, Neu5Gc has been found in the serum of cancer patients.[58] The melanomas exclusively overexpress a mannose transferase (DMP3) and a GlcNAc transferase that transfers GlcNAc exclusively to O-linked mannose (table 2). DMP3 is the dolichol mannose transferase responsible for transferring mannose to the mannosylated N-linked precursor. These two mannose transferases are part of two separate pathways, N-linked and O-linked. Their overexpression together may suggest a feed-forward regulatory loop based on the availability of mannose.

CONCLUSION

The cell lines share proliferation related genes such as the Notch genes and a few glycosyl transferases and glycosidases (Table A1). The shared glycan biosynthesis genes are mannosidase I (MANIA2), heparin sulfate 2-O-sulfotransferase 1 (HS2ST1), UDP-glucose ceramide glucosyltransferase like 2 (UGCGL2), polypeptidyl GalNAc transferase 10 (GALNT10), and carbohydrate sulfotransferase 10 (CHST10). No particular pathway is overrepresented. MANIA2 is associated with the N-linked glycosylation pathway. GALNT10 is one of the many ppGalNAc transferases in the O-linked glyocosylation pathway.

UGCGL2 is involved in ceramide synthesis in glycolipids. The lack of more of the N-linked glycosylation pathway genes might be related to the high cutoff for positives.

The overlap of positive signals of all genes for our data to previously published *Liu et al.* data for one cell line was only 25%. The percentage is not unusual given the previous debates on reproducibility of data, which is not an issue of technical reproducibility. In addition, immortalized cancer cell lines can have different growth properties dependent as they are cultured in different labs with different culture conditions.[59] It is also possible that the probe locations, or what part of the gene the probe targets, could affect overlap in the results.

The exon targeted by the probe also seems to vary in a tissue specific manner. This is true not only for MUC5B but for ABCF1 and CD44. The probes were designed from the 3′ end to avoid splice forms which normally occur from the 5′ end. There are few examples of 3′ splicing although it could be a possibility. Given that the mucins are highly duplicated in the genome, it might be interesting to consider the context of those genes within the genome to see whether they vary. Perhaps there are differences in enhancers at different promoter sights that lead to different copies being transcribed in different cell types and as the cell line ages. In the case of alternative splicing, splicing enhancers could also vary with cell type or age.

The differential expression of C-lectins and mucins in different cell types is quite interesting. C-lectins play a role in the homing of metastatic tumor cells to tissues.[60] The glycosylation of mucins has been difficult to study due to the complications in cleaving the O-glycans. Recently, using a new method, differences in the glycosylation of mucins in leukemia and epithelial cell lines

were found.[56] Our data shows that mucins are cell type specific suggesting that these mucins have specific roles. Whether that is determined at the glycan or transcript and how that is regulated is a growing field of research.

The cell type dependency of the probe location may explain the lack of correlation between gene calls and quantative PCR of glycogenes. Nairn et al. found a low correlation between the two, however or array suggests that the discrepancy was related to the number of positive gene calls. Positive genes are called by a platform dependent algorithm. The comparison was made with an Affymetrix array from the Consortium for Functional Glycomics which consisted of several short probes to different probe locations. Given that approximately 8% of our probe sets, which are sets of 3 highly specific probes to 1 exon, are cell line specific, our results illustrate the need for a glycogene array with probe locations at each exon.

METHODS

Clustering glycogenes from 2000 Ross Data

I downloaded genes with a 2-fold change in the Ross et al. data set from the Stanford Microarray Database, which is a generally accepted threshold for change in microarray experiments. Unfortunately, since the data was from an early gene microarray experiment, it was difficult to match the probe annotations. Many of them were EST tags which only referenced back to the array design without a RefSeq or Entrez gene ID, which would have made it easier to match to the list of glycogenes provided by the Consortium for Functional Glycomics. I decided to match partial terms related to glycosylation to the gene descriptions which were provided. For example, I used the search terms "glyco*," "mann*,""gluc*," where * is the wildcard designation. To ensure that I had exhausted the search terms, I checked my terms against the glycosylation relevant biosynthetic pathways in the KEGG database. The glycogene subset was hierarchically clustered using the Pearson centered correlation and average linkage

Removal of probes from the Nimblgen Catalog

Since the arrays were originally maximized for number of probes, we had to find the space for the new probes after they were designed. We began by using a published dataset and selecting genes that did not change and had a ratio between -0.5> x >0.5 for all of the NCI-60. We then manually selected 300 of the more than 1000 genes with low variance to remove from the Nimblegen gene catalog to make sure that we left in transcription factors or anything with a DNA

binding domain. We also removed a few of the random control probes which were in excess.

Culture of the NCI-60

The NCI-60 cells were cultured in RPMI (Cellgro) with 10% fetal bovine serum (Atlanta Biologicals) and 1% L-glutamine. Multiple people in the lab participated in growing the cells which were harvested at 80-90% confluency. For the RNA isolation, at least 5-10 million cells were trypsinized and stored in RNAlater (Qiagen) for up to 1 year. To minimize person to person variation, I performed the isolations, labeling and hybridization of sample to the expression arrays myself. The cells were isolated using the RNEasy mini-kit for total RNA isolation (Qiagen). I avoided precipitation of the RNA after isolation because the conditions can vary depending on the size of the RNA, which lead me to believe that there was some uncertainty regarding bias after precipitation. If the salt content was too high after the kit, I isolated the RNA from the stored stock again to insure the collection of a high quality data set. Occasionally, it was necessary to culture and harvested a different set of cells.

RNA quality evaluation

Successful RNA isolation was evaluated by spectrophotometry, using a Nanodrop. The relative absorbance of 280nm/260nm can be used to evaluate the presence of proteins, which have amino acids that absorb at 280nm. Ideally, this

ratio should be around 2, although a ratio of 1.8 is considered good. Similarly the presence of salts, which would disrupt downstream processing, is indicated by a high 230nm/260nm. The appropriate ratio is, once again, close to 2. If the total RNA isolation had two good ratios, the RNA quality was further tested using the Bioanalyzer, which is essentially a miniaturized electrophoresis gel. Using a combination of factors which includes the peak ratio for ribosomeal subunits, the Bioanalyzer calculates an RNA integrity number (RIN). Samples with an RIN of 8 or higher with a maximum value of 10 were used.

cDNA synthesis and labeling

The total RNA isolations were reverse transcribed with Superscript III (Invitrogen). Following the protocol in the Nimblegen expression guide, second strand synthesis was done using DNA Ligase and DNA Polymerase I (Invitrogen). The double stranded cDNA was labeled using a One-color labeling kit (Nimblegen). 12 of the cell lines were tested for good cDNA quality using the Bioanalyzer DNA 7500 kit (Agilent). All of them were good quality as evaluated by the specialists at the Genome Technology Center at the NYU Langone Medical Center.

Primer design

The primers were designed using NCBI primer blast. I input the RefSeq ID for the gene and selected the highest ranked primer pair (Fig. 3.5). These

primers were designed with a Tm = 70C and so that any unintended targets would have a 5 base mismatch. The primers were synthesized by IDT.

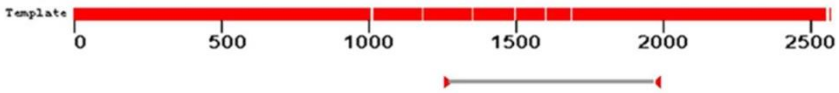
Rft1 NM_052859.3

	Sequence (5'->3')	Strand on template	Length	Start	Stop	Tm	GC%
Forward primer	AGGACTAGCCCTGCCGCCTG	Plus	20	3905	3924	60.32	70.00%
Reverse primer	GAGGCAGGCTGGCCCACATG	Minus	20	4087	4068	60.04	70.00%

CHST4 NM_005769.2

	Sequence (5'->3')	Strand on template	Length	Start	Stop	Tm	GC%
Forward primer	GCTTCCTCATTTGCTTCTCCCAGCC	Plus	25	68	92	60.06	56.00%
Reverse primer	AGCGCACAAGCAGGTAGCGTT	Minus	21	1037	1017	59.65	57.14%

St6GalNAc1 NM_018414.3



	Sequence (5'->3')	Strand on template	Length	Start	Stop	Tm	GC%
Forward primer	CGGTGCATCACCTGTGCCGT	Plus	20	1255	1274	59.97	65.00%
Reverse primer	GGCCCCGGTCAGTTCTTGGC	Minus	20	1985	1966	59.97	70.00%

Figure 3.5: Primer design for low copy glycogenes.

Gene array quality control

The data from the hybridized gene arrays was segmented in NimbleScan and the quality was evaluated according to their criteria for uniformity prior to inclusion in analyses resulting in 49 samples.

Chapter 4: Integration of Glycomics and Genomics

One challenge of interpreting microarray data is determining the biologically significant signals from the noise. As with any large data set, there is a tradeoff to setting signal thresholds which results in either false negatives or false positives. In addition, the usage of hierarchical clustering for classification has limitations. By definition, more similar sets of signals are grouped with each other before being compared and grouped to the next most similar set of signal. This process reduces the visibility of secondary patterns that might shared between two sets of signals that initially get separated into groups based on a primary pattern. Given that some of the glycosylation related genes have low copy number levels, reduction of noise without signal thresholding and sensitivity to lower signal patterns is of the utmost importance. Singular Value Decomposition (SVD) is a method that separates overlapping patterns into constituent metapatterns with an associated significance.[61] The focus can be shifted to the visualization of the significant patterns which could have biological or experimental relevance. Thus the signal is extracted from the noise independent of a threshold related to signal intensity. This chapter will discuss the application of SVD to the interpretation of the lectin array data from our lab and gene expression data published by another lab.

SINGULAR VALUE DECOMPOSITION : LECTIN ARRAY

Singular Value Decomposition (SVD) is a linear transformation of a data matrix into principal components (eigengenes and eigenarrays) with a diagonalized matrix of the contribution of each pattern. In its application to microarray data, the principal components represent a pattern across either all arrays or all genes[62]. Mathematically, this is represented as:

$$M = U\Sigma V^T$$

While similar to principal component analysis (PCA), the values in the Σ matrix facilitate the selection of most significant patterns. SVD has been used to study the yeast cell cycle and successfully identify novel genes involved in the process.[62] Generalized SVD works on similar mathematical principles and can be used to integrate matrices with a shared set of conditions[63].

The biological meaning of SVD can be difficult to interpret without recognizable patterns. In the case of the cell cycle data, the first two eigenarrays were sinusoidal, the hallmark of a cyclical process. With this in mind, we wanted to maximize our chances of seeing a pattern. To do this, we decided to focus on a subset of 4 cell types (melanoma, leukemia, colon and renal) which had distinct gene expression patterns in the Ross data. Each cell line was part of a dye swapped pair against a pooled reference that was hybridized to the lectin array. This pooled reference included one cell line from 6 different types that had a stable glycomic signature over multiple passages. The hierarchical cluster of the subset reveals some grouping by cell type which suggests a cell type

dependent glycosylation code (Fig. 4.1). In addition, it corroborated prior data from our lab wherein cells from different cell lines have clustered together.

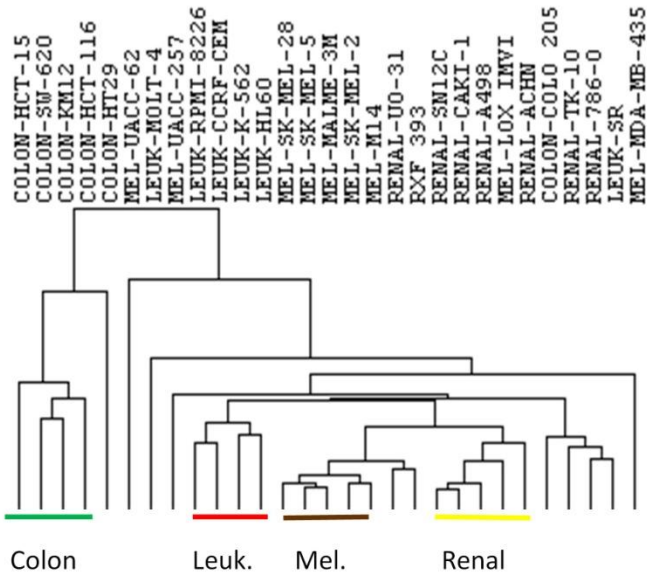


Figure 4.1: Heirarchical cluster of lectin array data for a subset of the NCI-60.

After a preliminary examination with hierarchical clustering, I decomposed the data using SVD (Fig. 4.2). For the SVD analysis, the cell lines were grouped by subset prior to decomposition. The resultant V matrix had 29 eigen-lectins which collectively describe patterns of lectin binding across the set of 29 cell lines.

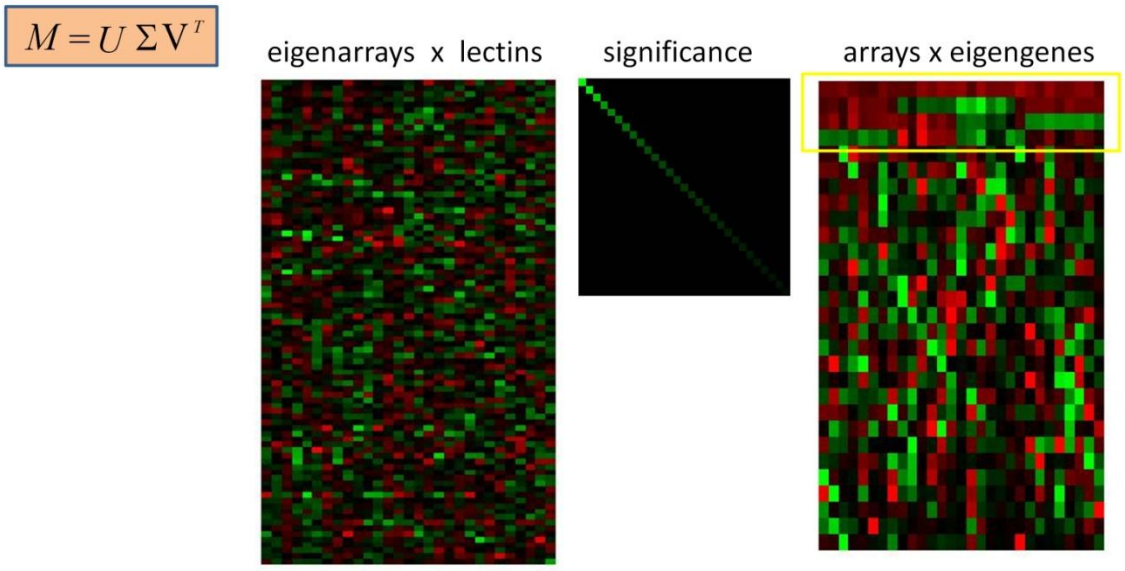


Figure 4.2: singular value decomposition of lectin array data for a subset of the NCI-60

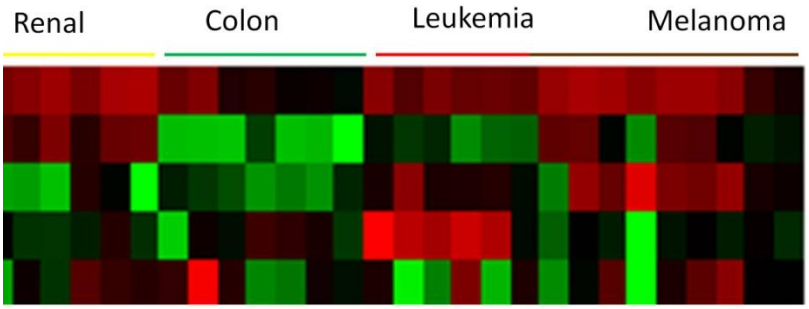


Figure 4.2: Close-up of the first 5 eigen-lectins in the SVD of the subset. The 2nd, 3rd, and 4th eigen-lectins show some cell type variation. The 5th is included for comparison.

The first eigen-lectin represents shared signals across the array (Fig. 4.2). Accordingly, little variation is observed. The second, third, and fourth eigen-lectins covary with cell type. The second eigen-lectin exposes a lectin binding pattern that is higher in the colon and leukemia cell lines, but lower in renal and melanoma.

The renal and colon cells are relatively higher than the leukemia and melanoma cells for eigen-lectin 3. The overall pattern for eigen-lectin 4 is hard to interpret, but there the signals are clearly lower for leukemia. The eigen-lectins are ordered by significance. While some techniques have been developed to threshold eigen-patterns based on their significance which can be calculated from the diagonalized Σ matrix, the cell type dependent eigen-lectins are amongst the top and their relative significance can be presumed.[64] Moreover, their clear cell type dependent pattern suggests biological importance.

INTEGRATION OF A SUBSET

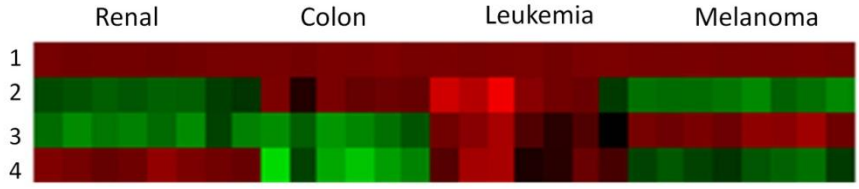
With this preliminary data in hand, we decided to integrate the lectin array data with a high quality published dataset.[65] This dataset was released while we were collecting our own and includes both mRNA and miRNA isolations hybridized to and Agilent platform. The whole genome expression array data previously available for the NCI-60 was hybridized to a non-commercial microarray which would have made annotation matching extremely laborious.[54] The Agilent platform is widely utilized and the probe IDs have been deposited in a number of databases including DAVID. DAVID is a resource provided by the NCBI which houses a number of different resources including an ID converter.[66,67] Agilent probe IDs can easily be converted to Entrez gene IDs.

To integrate the data we used a generalized version of SVD (GSVD) (Fig. 4.3a). This method has been successfully used in the integration of yeast and human microarray data.[68] GSVD integratation of two distinct datasets also illustrates a key difference between principal component analysis and singular value decomposition. Principal Component Analysis decomposes the data into two matrices and lacks the Σ matrix, which is not only useful for calculating significance but a scaling matrix.[61] Without it, two datasets on different scales could not be compared well. Even with a ratiometric approach, the data from two different microarrays could be on different scales, which is even more probable for microarrays of different types, such as the lectin and gene arrays.

For the GSVD integration, the data was again organized by cell type. The resultant matrix of eigen-gene-lectins and arrays has 3 cell type dependent eigen-

gene-lectins (called eigen-celltype from this point onward) which appear to be better defined than the eigen-lectins in the SVD of the lectins alone. I projected the vector of a lectin signal for all cell types onto the eigen-celltypes to calculate the correlation of each lectin signature to the eigen-celltype pattern (Fig. 4.3b).[69] This is measure of the gene or lectin similarity to the cell type pattern. I filtered the correlation matrix for genes and lectins that were exclusively positive to one of the 3 eigen-celltypes when compared to the first 4. I chose a stringent filter because signals could poorly correlate to multiple eigen-celltypes and many of the lectins have overlapping specificities.

a.



b.

Eigen-celltype 2	APA, Con A, DSA , GNA, LcH, LPA, NPA, PAA ,SNA (vector), UDA, HHL, MAL-II, CVN, SVN, GRFT, ACA, PSL, TJA-I, AMA, Calsepa, IRA, RS-IIL	High mannose, 2,6 sialic acid (n-link)
Eigen-celltype 3	VRA, PTL-II, HAA	Terminal GalNAc (o-link)
Eigen-celltype 4	TKA, PapGII	Glycolipid

Figure 4.3: A) Close-up of the first for eigen-celltypes in the GSVD of both the lectins and the genes. B) Projection correlation was calculated for the lectins. Lectins exclusive to an eigen-lectin are summarized.

The lectins which correlated to the second eigen-celltype and therefore were higher in the renal and melanoma cell lines were lectins that bind to N-glycans. More specifically, there is a strong high mannose signature (CVN, SVN, GRFT, Calsepa, AMA, NPA, UDA, HHL) and α2,6 sialic acid (SNA, PSL, TJA-I). Experiments in our lab using the melanoma cells have separately observed a high mannose signal. In the third eigen-celltype, the lectins that were exclusive to that eigen-celltype were terminal GalNAc specific (VRA, PTL-II, HAA) which implies O-linked glycans. Although there are only 3 exclusive lectins, a less stringent filter to include any positively correlated lectin includes more O-linked specific lectins. This eigen-celltype is higher in the renal and colon cells. The fourth eigen-celltype is higher in glycolipid binding lectins (TKA, PapGII) and higher in the colon and melanoma cell lines.

Enrichment Analysis

The genes were treated the same way as the lectins. They were projected onto the eigen-celltypes and the genes that were exclusively correlated to one eigen-celltype were examined further. Looking for enrichment of annotations associated with a biological pathway is an easy and informative way to study large data sets. The gene list was uploaded to the DAVID database and their ontologies were retrieved. I chose DAVID, rather than another gene ontology database, because it displays the known shared ontologies of clustered genes in a matrix format. This is particularly important for glycogenes which are not always listed as the same process in gene ontology databases, for example in the case of sialic acid.[70] The enriched gene ontology clusters have an enrichment score, which allows you rank the clusters by significance although other clusters

could still be interesting. The cluster enrichment scores are the mean of the P-values of all the annotations in the cluster. The P-values are calculated from the percentage of annotations of a particular biological pathway observed in the gene set compared to the percentage of those annotations that would be expected from the human genome.

Taking the top 3 gene clusters from each set, I was able to observe some interesting trends (Table 4.1).

	Eigen-celltype 2 (N-glycan)	Eigen-celltype 3 (O-glycan)	Eigen-celltype 4 (glycolipid)
Cluster 1	Ribonucleotide binding (score = 6.41)	Organelle lumen (score = 55.83)	Mitochondrial (score = 24.92)
Cluster 2	Cytoskeletal, microtubule(score = 5.48)	Ribosomal (score = 45.55)	Organelle lumen (score = 23.54)
Cluster 3	Organelle lumen (score = 5.03)	mRNA splicing and processing (score =42.4)	Mitochondrial (score =18.11)

Table 4.1: Summary of top 3 gene ontology clusters in DAVID for each eigen-celltype.

Genes related to organelle lumen are highly expressed in all three eigen-celltypes. The biosynthetic pathway of glycosylation occurs in the secretory pathway, through the endoplasmic reticulum and golgi, so it is encouraging organellar lumen classified genes are being expressed.

As I said previously, each eigen-celltype corresponded to a different family of glycosylation. Interestingly, the most significant gene ontology clusters were from distinct pathways. Genes involved in cytoskeletal and microtubule regulation, as well as ribonucleotide binding proteins, were enriched in the same eigen-celltype that correlated to N-glycosylation. However, the enrichment score for the clusters in eigen-celltype 2 are low compared to eigen-celltype 3 and 4. The highest enrichment score in that set is 4.09 compared to enrichment scores with maximum values of 55.83 and 24.92, for eigen-celltypes 3 and 4. Since eigen-celltype 2 is has a higher significance than eigen-celltypes 3 and 4, it represents a more general process, which might explain the low enrichment scores. This suggests less influence from other cellular processes for N-glycosylation or that the regulators are too few to be clustered in the ontology enrichment. As previously introduced, the N-glycosylation pathway is highly conserved and can be modeled as a modular synthetic route through the secretory pathway. In addition, studies of N-glycosyslation have already shown a high correlation with transcript level and glycan structure. So the N-glycosylation pathway appears to be well defined and for the most part, regulated at the transcript level of the genes in the pathway.

The gene list for this eigen-celltype includes several genes that would be expected with N-glycosylation, such as class II alpha mannosidases which are involved in trimming the tri-mannose N-glycosylation core. The function of the class II mannosidases is not as well known as the class I mannosidases which signal misfolded proteins for degradation.[71] α-mannosidase II is the second mannosidase in the N-glycosylation pathway which cleaves mannose on the third branch. If the mannose is not cleaved, the branch is not elaborated further,

resulting in a partially elaborated and stunted hybrid glycan. Inhibition of α-mannosidase II in HEK293 cells results in hybrid N-glycans.[72]

Eigen-celltype 3 corresponded to o-glycosylation and was enriched in ribosomal and RNA processing genes. The regulation of O-glycosylation is not well known. The first step in O-glycosylation is the transfer of a UDP-GalNAc with one of over 20 polypeptide galnac transferases.[73] These transferases can compensate for each other making knock-out studies difficult to interpret. Currently, little is known of the required sequence for O-glycosylation, in part, this is due to the difficulty of cleaving O-glycans in an unbiased way from a protein. Some differential binding of groups of the polypeptide Galnac transferases has been demonstrated. However, the requirement is only a basic residue two residues from the serine or threonine suggesting promiscuity and redundancy. Recently, evidence of COP-I retrograde translocation of the polypeptide GalNAc transferases from the Golgi the trans endoplasmic reticulum upon EGF activation was reported. EGF activation resulted in higher overall amounts of o-glycosylation.[74] The promiscuity and redundancy of the polypeptide galnac transferases could be regulated through trafficking which might change in dynamic circumstances. This type of system could rapidly generate variability which could modulate signals. It has been suggested that carbohydrates are participate in "analog" type signaling.

Intriguingly, a cluster of genes with O-GlcNAc transferase (OGT) and 3 different TPR domains were identified within the same eigen-celltype. O-GlcNAc is a small glycan involved in signaling whose roles have not been completely elucidated. It is thought that the TPR domains participate in substrate selection and they themselves are substrates for OGT.[75,76] O-GlcNAc is

required for the aggregation of untranslated ribosomes to stress bodies so it is interesting that it coexpresses with translational machinery in the eigen-celltype which is accociated with O-glycosylation.[77]

The lectins that correlated to the 4th eigen-celltype represented glycosylated lipids. Mitochondrial genes were the top gene clusters for this eigen-celltype. Mitochondrial associated membranes (MAM) which bridge mitochondria to the ER have been associated with lipid synthesis and trafficking of lipids during stress.[78] Evidence of a glucosyl and GalNAc transferase in the MAM has been demonstrated.[79] The mitochondria are highly influenced by metabolism and this data suggests that might also influence lipid glycosylation. Perhaps, this offers an explanation for the increased GalNAc in breast cancer, which is visible by HPA binding.

The patterns that were decomposed in the SVD of the lectins are actually different from those in the GSVD of the lectins and genes together (Fig. 4.4a). The second eigenvectors are reciprocal patterns. The third eigenvectors are similar in pattern. The fourth set is mixed with one half of the signals in reciprocity and the other in concordance. Biplots of the projection correlation of each lectin onto the eigenvectors from the decomposition of lectin signals alone and with the gene signals provide a better sense of the trend. The third eigenvectors are highly positively correlated to each other as expected (Fig. 4.4c). Surprisingly, the exclusive lectins do not overlap as lectins although their specificity for o-glycans do to some extent. HAA,PTL-II,MNA-G, MPA are GalNAc binding lectins. TJA-II and VRA are galactose binding lectins. The other lectins are a combination of glycolipid and N-glycan lectins. The second eigenvectors are negatively correlated in the fourth quadrant which confirms the

patterns visible in the heatmaps (Figure 4.4a). The most extreme fourth quadrant lectins have a positive correlation to the eigen-celltype which more closely reflects the gene expression and a negative correlation to the eigen-lectin which reflects the lectin binding or carbohydrate expression. Suggestively, at this extremity, the lectins largely correspond to high mannose lectins (HHL, SVN, GRFT, AMA, TL, UDA). This could signify that there is high gene regulation of the high mannose glycans in renal and melanoma cells.

No trend in terms of the eigen-celltype versus eigen-lectin biplot is observable for the fourth eigenvectors (4-4d). Perhaps, the relationship with metabolism is influencing this distribution. It will be interesting to see whether the breast cell lines are described with this signature, which did not cluster as a cell type in the Ross data.

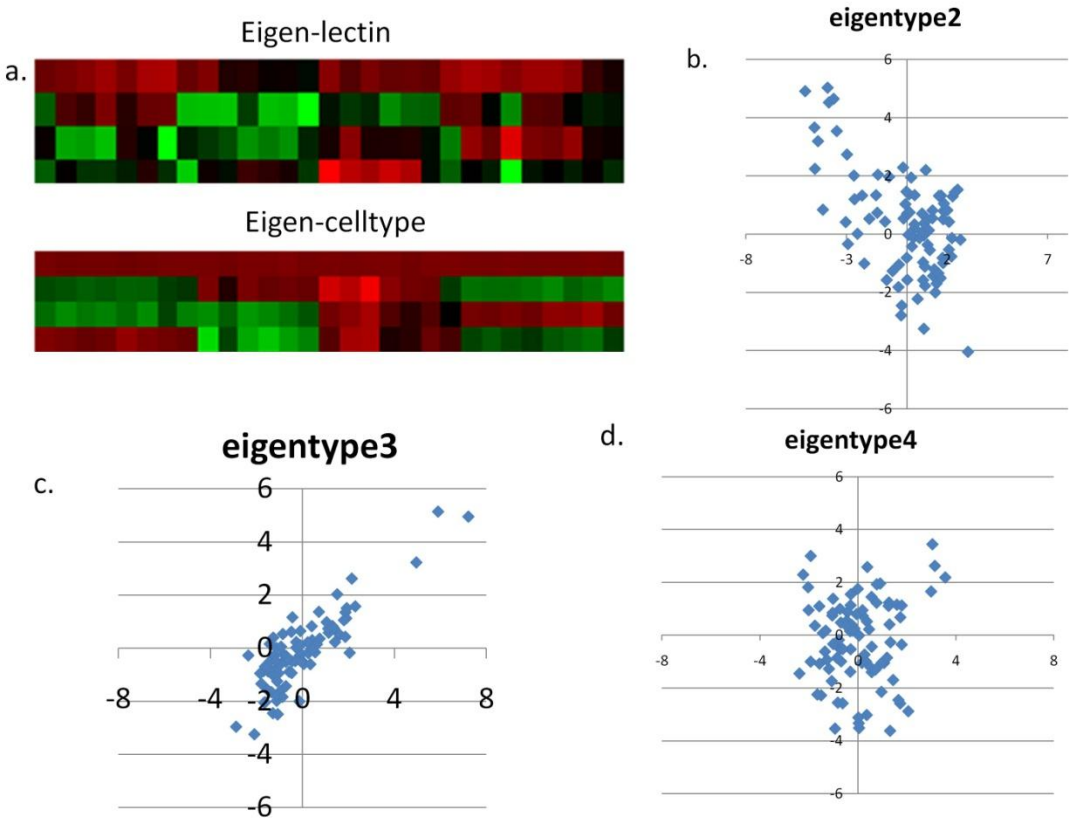


Figure 4.4: Comparison of SVD of lectin signals and GSVD of gene array signals using biplots.

CONCLUSIONS

The biplots from the SVD of the lectins and the GSVD of the lectins and genes suggest that the regulation of N- and O-glycosylation which differentiates these cell lines is different. In the case of N-glycosylation, the biplot has a strong negative trend which is also clearly visible as a reciprocal relationship in the SVD heatmaps. This suggests pivotal change which would be the case in binary decision that is simply regulated by the level of the enzyme. Given that the N-glycosylation pathway is highly conserved, regulated at the transcript level and sequential, this model makes sense. We find that binding of lectins with a specifity to complex N-glycan structure are inversely correlated to α-mannosidase expression. This is expected given the role of α-mannosidase II in cleaving a mannose residue from one of the core branches prior to elaboration, which leads to the complex N-glycan type. The high mannose binding proteins correlated with the eigen-celltype and α-mannosidase II. A-mannosidase II levels determine the fate of N-glycans to either a complex or high mannose type. This corresponds to what is generally known for the mannosidases which in terms of N-linked glycan elaboration is early in the pathway. In addition, the α2,6 sialic acid lectins (SNA, PSL, TJA-I) correlate to the eigen-celltype, but the lectins correlated to the eigen-lectin were both α2,6 and α2,3 sialic acid binders (SNA, MAA). The correlated gene list only includes α2,3 sialic acid transferase and not α2,6. This suggests some inverse correlation between the α2,6 and α2,3 sialic acid transferases. It is unlikely to be non-specific binding of α2,6 sialic acid lectins because 3 of them were correlated to the eigen-celltype. Sialic acid addition is one of the final steps in the N-linked glycan elaboration. A previous

study on N-glycosylation glycomics reported low correlation of transcript levels for both oligomannose and sialic acid. Our integration is more informative and shows and inverse correlation between the genes regulating oligomannose and their cell surface display. In addition, our data suggests that a similar relationship may exist for the α2,3 and α2,6 sialic acids.

It would be interesting to consider these trends within the context of the other correlated genes. Although, the cytoskeletal genes are not as highly enriched, there may be a few key regulators. During cell migration, there is dynamic feedback between adhesion protein and the cytoskeleton.[80] E-cadherin is a well studied adhesion molecule with roles in cancer. Its dysfunction cannot be fully explained with genetics or epigentics. Changes in N-glycosylation have been implicated with its dysfunction in some carcinomas.[2] Although, E-cadherin was not in our genelist, it is tempting to speculate on the roles of altered N-glycans in adhesion molecules.

The O-glycosylation SVD biplot is linear and positively correlated which is promising in terms of understanding the regulation of O-glycosylation. In light of the fact that the gene ontology clusters have high enrichment scores, these genes may actually play a large role in the O-glycosylation. Some of the genes in the gene list were expected and confirm that the correlation and decomposition was successful. Two polypeptide GalNAc transferases, 2 and 7, correlate to the eigen-celltype, although strangely, none of the mucins do. Our own gene array data suggests cell type specific mucins and some exon specificity. In fact, the mucin signals we observed were all from the custom designed probes. Given the redundancy of the polypeptide GalNAc transferases and that the enriched ontologies suggest a role for translation machinery, it may

be difficult to make further hypotheses without perturbing the cell. In addition, this data set lacks adequate mucin probes. In the future, it could be interesting to investigate the regulation of alternative splicing of mucins and the relationship of mRNA degradation pathways. In addition, comparing the regulation of mucins compared to the regulation of polypeptidyl transferases in creating functionally diverse and cell or possibly tissue specific mucins would also be interesting.

METHODS

The SVD and GSVD were performed Mathematic (Wolfram) using the built-in function. The projections were also calculated in Mathematic by taking the dot product of the eigen-celltype or eigen-lectin with every lectin and gene. The data was filtered in Excel using conditional statements and auto-filter. The genelists were uploaded to DAVID (http://david.abcc.ncifcrf.gov/) for the ontology enrichment.

Conclusions

The data suggests that these four cell types have a glycosylation signature and suggests that other cell types do as well. Our data also shows that the glycosylation that differentiates each cell type can be decomposed into the major glycosylation families, N-linked, O-linked and glycolipid. Based on gene ontology enrichment and biplots of the correlation of each lectin to the eigen-lectins and eigen-celltypes, it appears that that the glycosylation that differentiates these cell types is regulated by different mechanism dependent on the glycosylation type (N-,O-, glycolipid). Although these cells are cancerous, they still reveal differences in the regulation of cell type dependent glycosylation. In addition, our ability to decompose the data using previously published work suggests that these patterns are very strong and that integration with our data set may reveal more subtleties. Extension of the integration to other cell lines and using the customized gene array data with more glycosylation probes is planned. We may encounter some problems with cell types that are more diverse, such as the breast cell lines. However, we could investigate some of these hypotheses in larger, publicly available datasets. Genome wide association studies have illustrated the functional impact of mutations in polysialic acid transferase. Our data suggests cell type specific exons for mucins. This implies the need for a larger cohort which might be publically available. Alternatively, we could look for glycogenes in the copy number variance data for the NCI-60 or sequence our own cell lines. The integration of multiple environmental and hereditary factor in the misregulation of the N-glycosylation pathway in the Golgi which explains

the pathology of multiple sclerosis is a poignant example of the importance and adaptability of glycosylation pathways.[5] Our data suggests that the regulation of each type of glycosylation pathway, particularly for N- and O- links, is different. In the future, we hope to investigate the role of other elements in each pathway and any differences that may suggest a difference in the evolution of glycosylation pathways. Other master regulators may be identified and perturbed. Thus, we hope to move toward a model that sufficiently describes the regulation of glycosylation in the context of the dynamics in the secretory pathway.

Appendix

Table A1: Shared glycogenes

Glycogene Pathway	**Glycogene Family**	**Name**
CBP:C-Type Lectin	Novel	CD83 antigen
CBP:C-Type Lectin	5-NK Receptors	CLEC12B (macrophage antigen h)
CBP:I-Type lectin	Siglec	MAG [SIGLEC4A]
CBP:I-Type lectin	Siglec	MAG [SIGLEC4A]
Galectin	Galectin	LGALS9 (Galectin 9)
Glycan Degradation	Lysozomal Enzymes/Proteins	AGA (aspartylglucosaminidase precursor)
Glycan Degradation	Mannosidase	MAN1A2 [mannosidase alpha class 1A member 2]
Glycan-transferase	Sulfo-T	HS2ST1 [heparan sulfate 2-O-sulfotransferase 1]
Glycan-transferase	Glc-T	UGCGL2 [UDP-glucose ceramide glucosyltransferase-like 2]
Glycan-transferase	GalNAc-T	GALNT10 (ppGalNAc T10; GalNAc transferase 10)

Glycan-transferase	Sulfo-T	CHST10 [carbohydrate sulfotransferase 10]
Glycoproteins	Mucins	MUC1 [Mucin 1, cell surface associated]
Glycoproteins	Mucins	EMR2 [egf-like module containing, mucin-like, hormone 2]
Glycoproteins	Mucins	MUC1 [Mucin 1, cell surface associated]
Glycoproteins	Mucins	MUC5B [mucin 5 subtype B tracheobronchial]
intracellular protein transport	Golgi tethering factor	COG5 [component of oligomeric golgi complex 5]
intracellular protein transport	Golgi tethering factor	COG5 [component of oligomeric golgi complex 5]
Notch pathway	Notch Ligands	JAG2 [jagged 2]
Notch pathway	Notch Receptors	Notch2 (variant)
Notch pathway	Notch Receptors	Notch4 (variant)
Nuc. Sugar	Nucleotide Synthesis	GALE [UDP-galactose-4-epimerase]
Nuc. Sugar	Nuc. Sugars Transporters	SLC35E4 [solute carrier family 35 member E4]
xChemokine	MCP	CCL2 [small inducible cytokine A2]
xGrowth Factors & Receptors	FGF&R	FGFR3 [fibroblast growth factor receptor 3]

xGrowth Factors & Receptors	Miscellaneous	FGFR1 [fibroblast growth factor receptor 1]
xGrowth Factors & Receptors	HGF	HGF [hepatocyte growth factor isoform 1 or 3]
xHuman Housekeeping	xHuman Housekeeping	ABCF1 [ATP-binding cassette, sub-family F (GCN20), member 1]
xHuman Housekeeping	xHuman Housekeeping	SEPT2 [septin 2]
xHuman Housekeeping	xHuman Housekeeping	HNRPD [heterogeneous nuclear ribonucleoprotein D (AU-rich element RNA binding protein 1, 37kDa]
xInterleukin & Receptors	IL	IL16 [interleukin 16]
xInterleukin & Receptors	IL receptor	IL6R [interleukin 6 receptor]
xInterleukin & Receptors	IL	IL1R2 [interleukin 1 receptor, type II]
xMiscellaneous	Miscellaneous	EIF4G2 [eukaryotic translation initiation factor 4]
xMiscellaneous	Miscellaneous	EIF3C [eukaryotic translation initiation factor 3]
xMiscellaneous	Miscellaneous	HSP90AA1 [heat shock protein 90kDa alpha (cytosolic)]
xMiscellaneous	Miscellaneous	HDLBP [high density lipoprotein binding protein]
xProteoglycan	Miscellaneous	CD74
xSulfotransferas	Sulfo-T Protein	TPST2 [tyrosylprotein sulfotransferase 2]

e	tyrosine	
xSulfotransferase	Sulfo-T Protein tyrosine	TPST2 [tyrosylprotein sulfotransferase 2]
xSulfotransferase	Sulfo-T Protein tyrosine	TPST1 [tyrosylprotein sulfotransferase 1]

Table A2: Cell type specific glycogenes

CBP:C-Type Lectin	4-Selectin	SELL [selectin L precursor]
CBP:C-Type Lectin	5-NK Receptors	CLEC7A (Dectin-1)
CBP:C-Type Lectin	6-MMR	MRC1L1 or MRC1
CBP:C-Type Lectin	6-MMR	PLA2R1 (Phospholipase A2 receptor 1)
CBP:I-Type Lectin	Non-Siglec	N-CAM CD56
CBP:I-Type lectin	Siglec	MAG [SIGLEC4A]
Glycan-transferase	GlcNAc-T	POMGNT1 (related transcript: RP11-322N21)

Glycan-transferase	Man-T	DPM3
Glycoproteins	Mucins	MUC17 [Mucin 17, cell surface associated]
Glycoproteins	Mucins	MUC6 [mucin 6 gastric]
Glycoproteins	Mucins	MUC5B [mucin 5 subtype B tracheobronchial]
Glycoproteins	Mucins	MUC5B [mucin 5 subtype B tracheobronchial]
Glycoproteins	Mucins	MUC5AC
intracellular protein transport	Golgi tethering factor	COG5 [component of oligomeric golgi complex 5]
intracellular protein transport	Golgi tethering factor	COG3 [component of golgi transport complex 3]
Notch pathway	Notch Receptors	Notch3 - Short Trans
Nuc. Sugar	Nucleotide Synthesis	CMP-NeuAc hydroxylase-like protein [Cytidine monophosphate-N-acetylneuraminic acid hydroxylase-like protein]
Nuc. Sugar	Nuc. Sugars Transporters	SLC35C2 [solute carrier family 35 member C2]
Nuc. Sugar	Nucleotide Synthesis	GALE [UDP-galactose-4-epimerase]
xAdhesion Molecule	Adhesion Molecule	CD48 [CD48 molecule]

xCytokine	Miscellaneous	TCEA1 [transcription elongation factor A 1]
xGrowth Factors & Receptors	Miscellaneous	FGFR1 [fibroblast growth factor receptor 1]
xGrowth Factors & Receptors	VEGF	VEGF [vascular endothelial growth factor A]
xGrowth Factors & Receptors	FGF&R	FGFR2 [fibroblast growth factor receptor 2]
xHuman Housekeeping	xHuman Housekeeping	ABCF1 [ATP-binding cassette, sub-family F (GCN20), member 1]
xMiscellaneous	Miscellaneous	EIF3C [eukaryotic translation initiation factor 3]
xProteoglycan	Miscellaneous	CD44 (Epican)
xProteoglycan	Miscellaneous	CD44 (Epican)

References

1. Haltiwanger, R.S. & Lowe, J.B. Role of glycosylation in development. *Annual review of biochemistry* **73**, 491-537(2004).

2. Rambaruth, N.D.S. & Dwek, M.V. Cell surface glycan-lectin interactions in tumor metastasis. *Acta histochemica* **113**, 591-600(2011).

3. Nothaft, H. & Szymanski, C.M. Protein glycosylation in bacteria: sweeter than ever. *Nature reviews. Microbiology* **8**, 765-78(2010).

4. Rakus, J.F. & Mahal, L.K. New technologies for glycomic analysis: toward a systematic understanding of the glycome. *Annual review of analytical chemistry (Palo Alto, Calif.)* **4**, 367-92(2011).

5. Mkhikian, H. et al. dysregulate N-glycosylation in multiple sclerosis. *Nature Communications* **2**, 334-13(2011).

6. Isomura, R., Kitajima, K. & Sato, C. Structural and Functional Impairments of Polysialic Acid by a Mutated Polysialyltransferase Found in Schizophrenia * □. *Journal of Biological Chemistry* **286**, 21535-21545(2011).

7. Stevens, J. et al. influenza viruses. *Microbiology* **4**, 857-864(2006).

8. Kim, P.-J., Lee, D.-Y. & Jeong, H. Centralized modularity of N-linked glycosylation pathways in mammalian cells. *PloS one* **4**, e7317(2009).

9. Gill, D.J. et al. Regulation of O-glycosylation through Golgi-to-ER relocation of initiation enzymes. *The Journal of cell biology* **189**, 843-58(2010).

10. Metallo, C.M. & Heiden, M.G.V. Metabolism strikes back : metabolic flux regulates cell signaling. *Genes & Development* 2717-2722(2010).doi:10.1101/gad.2010510.feedback

11. Cummings, R.D. The repertoire of glycan determinants in the human glycome. *Molecular bioSystems* **5**, 1087-104(2009).

12. Cantarel, B.L. et al. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic acids research* **37**, D233-8(2009).

13. Varki, A. et al. *The Essentials of Glycobiology*. (2009).

14. Hashimoto, K. et al. Comprehensive analysis of glycosyltransferases in eukaryotic genomes for structural and functional characterization of glycans. *Carbohydrate research* **344**, 881-7(2009).

15. Song, X. et al. Quantifiable fluorescent glycan microarrays. *Glycoconjugate journal* **25**, 15-25(2008).

16. Pilobello, K.T. et al. Development of a lectin microarray for the rapid analysis of protein glycopatterns. *Chembiochem : a European journal of chemical biology* **6**, 985-9(2005).

17. Zaia, J. Mass spectrometry and glycomics. *Omics : a journal of integrative biology* **14**, 401-18(2010).

18. Ebe, Y. et al. Application of lectin microarray to crude samples: differential glycan profiling of lec mutants. *Journal of biochemistry* **139**, 323-7(2006).

19. Oyelaran, O. & Gildersleeve, J.C. Glycan arrays: recent advances and future challenges. *Current opinion in chemical biology* **13**, 406-13(2009).

20. Gabius, H.-J. et al. The sugar code: functional lectinomics. *Biochimica et biophysica acta* **1572**, 165-77(2002).

21. Comelli, E.M. et al. A focused microarray approach to functional glycomics: transcriptional regulation of the glycome. *Glycobiology* **16**, 117-31(2006).

22. North, S.J. et al. Glycomics profiling of Chinese hamster ovary cell glycosylation mutants reveals N-glycans of a novel size and complexity. *The Journal of biological chemistry* **285**, 5759-75(2010).

23. Nairn, A.V. et al. Regulation of glycan structures in animal tissues: transcript profiling of glycan-related genes. *The Journal of biological chemistry* **283**, 17298-313(2008).

24. Yamamoto-Hino, M. et al. Identification of genes required for neural-specific glycosylation using functional genomics. *PLoS genetics* **6**, e1001254(2010).

25. Hizukuri, Y. et al. Extraction of leukemia specific glycan motifs in humans by computational glycomics. *Carbohydrate research* **340**, 2270-8(2005).

26. Mammen, M., Choi, S.-K. & Whitesides, G.M. Polyvalent Interactions in Biological Systems: Implications for Design and Use of Multivalent Ligands and Inhibitors. *Angewandte Chemie International Edition* **37**, 2754-2794(1998).

27. Rüdiger, H. & Gabius, H.J. Plant lectins: occurrence, biochemistry, functions and applications. *Glycoconjugate journal* **18**, 589-613(2001).

28. Yamamoto, K. et al. Measurement of the carbohydrate-binding specificity of lectins by a multiplexed bead-based flow cytometric assay. *Analytical biochemistry* **336**, 28-38(2005).

29. Brooks, S. a, Hall, D.M. & Buley, I. GalNAc glycoprotein expression by breast cell lines, primary breast cancer and normal breast epithelial membrane. *British journal of cancer* **85**, 1014-22(2001).

30. Fry, S. a et al. Lectin microarray profiling of metastatic breast cancers. *Glycobiology* **21**, 1060-70(2011).

31. Krishnamoorthy, L. et al. arguing for a common origin. *October* **5**, 244-250(2009).

32. Schena, M. et al. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science (New York, N.Y.)* **270**, 467-70(1995).

33. Rao, Y. et al. A comparison of normalization techniques for microRNA microarray data. *Statistical applications in genetics and molecular biology* **7**, Article22(2008).

34. Quackenbush, J. Microarray data normalization and transformation. *Nature genetics* **32 Suppl**, 496-501(2002).

35. Pilobello, K.T., Slawek, D.E. & Mahal, L.K. A ratiometric lectin microarray approach to analysis of the dynamic mammalian glycome. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 11534-9(2007).

36. Tran, P.H. et al. Microarray optimizations: increasing spot accuracy and automated identification of true microarray signals. *Nucleic acids research* **30**, e54(2002).

37. Li, Q. et al. Donuts, scratches and blanks: robust model-based segmentation of microarray images. *Bioinformatics (Oxford, England)* **21**, 2875-82(2005).

38. Do, J.H. & Choi, D.-kug Minireview Molecules and Normalization of Microarray Data : Single-labeled and Dual-labeled Arrays. **22**, 254-261(2006).

39. Mary-Huard, T. et al. Statistical methodology for the analysis of dye-switch microarray experiments. *BMC bioinformatics* **9**, 98(2008).

40. Dudoit, S. et al. STATISTICAL METHODS FOR IDENTIFYING DIFFERENTIALLY EXPRESSED GENES IN REPLICATED cDNA MICROARRAY EXPERIMENTS. *Statistica Sinica* **12**, 111-139(2002).

41. Lu, R. et al. Assessing probe-specific dye and slide biases in two-color microarray data. *BMC bioinformatics* **9**, 314(2008).

42. Smyth, G.K. & Speed, T. Normalization of cDNA microarray data. *Methods* **31**, 265-273(2003).

43. Wang, X. et al. A novel approach for high-quality microarray processing using third-dye array visualization technology. *IEEE transactions on nanobioscience* **2**, 193-201(2003).

44. Otzen, D. Protein-surfactant interactions: a tale of many states. *Biochimica et biophysica acta* **1814**, 562-91(2011).

45. Propheter, D.C., Hsu, K.-L. & Mahal, L.K. Fabrication of an oriented lectin microarray. *Chembiochem : a European journal of chemical biology* **11**, 1203-7(2010).

46. Propheter, D.C. & Mahal, L.K. Orientation of GST-tagged lectins via in situ surface modification to create an expanded lectin microarray for glycomic analysis. *Molecular bioSystems* **7**, 2114-7(2011).

47. Eisen, M.B. et al. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 14863-8(1998).

48. Assistance, F.T. & Solutions, M. Nexterion ® Slide H. (2009).

49. Sweryda-Krawiec, B. et al. A new interpretation of serum albumin surface passivation. *Langmuir : the ACS journal of surfaces and colloids* **20**, 2054-6(2004).

50. Haferlach, T. et al. Clinical utility of microarray-based gene expression profiling in the diagnosis and subclassification of leukemia: report from the International Microarray Innovations in Leukemia Study Group. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **28**, 2529-37(2010).

51. Shi, L. et al. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature biotechnology* **24**, 1151-61(2006).

52. Birney, E. et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799-816(2007).

53. Comelli, E.M. et al. A focused microarray approach to functional glycomics: transcriptional regulation of the glycome. *Glycobiology* **16**, 117-31(2006).

54. Ross, D.T. et al. Systematic variation in gene expression patterns in human cancer cell lines. *Nature genetics* **24**, 227-35(2000).

55. Pruitt, K. Chapter 18 : The Reference Sequence (RefSeq) Database Database Content : Background.

56. Lesuffleur, T. et al. Differential expression of the human mucin genes MUC1 to MUC5 in relation to growth and differentiation of different mucus-secreting HT-29 cell subpopulations. *Journal of cell science* **106 (Pt 3**, 771-83(1993).

57. Varki, A. Loss of N-Glycolylneuraminic Acid in Humans : Mechanisms , Consequences , and Implications for. *Yearbook of Physical Anthropology* **69**, 54 - 69(2001).

58. Inoue, S., Sato, C. & Kitajima, K. Extensive enrichment of N-glycolylneuraminic acid in extracellular sialoglycoproteins abundantly synthesized and secreted by human cancer cells. *Glycobiology* **20**, 752-62(2010).

59. Irelan, J.T. et al. Rapid and quantitative assessment of cell quality, identity, and functionality for cell-based assays using real-time cellular analysis. *Journal of biomolecular screening* **16**, 313-22(2011).

60. Drickamer, K. & Taylor, M.E. Minireview Glycan arrays for functional glycomics. *Genome* 10-13(2002).

61. Practical, I.A. et al. Chapter 5 Singular value decomposition and principal component analysis. 1-18(2003).

62. Alter, O., Brown, P.O. & Botstein, D. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 10101-6(2000).

63. Alter, O., Brown, P.O. & Botstein, D. Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 3351-6(2003).

64. Epps, B.P. & Techet, A.H. An error threshold criterion for singular value decomposition modes extracted from PIV data. *Experiments in Fluids* **48**, 355-367(2009).

65. Liu, H. et al. mRNA and microRNA Expression Profiles of the NCI-60 Integrated with Drug Activities. *Molecular cancer therapeutics* **9**, 1080-1091(2010).

66. Huang, D.W., Sherman, B.T. & Lempicki, R. a Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols* **4**, 44-57(2009).

67. Huang, D.W., Sherman, B.T. & Lempicki, R. a Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research* **37**, 1-13(2009).

68. Alter, O., Brown, P.O. & Botstein, D. Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 3351-6(2003).

69. Alter, O. & Golub, G.H. Integrative analysis of genome-scale data by using pseudoinverse projection predicts novel correlation between DNA replication and RNA transcription. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 16577-82(2004).

70. Cohen, M. & Varki, A. The Sialome — Far More Than the Sum of Its Parts. *OMICS: A Journal of Integrative Biology* **14**, (2010).

71. Helenius, A. & Aebi, M. Roles of N-linked glycans in the endoplasmic reticulum. *Annual review of biochemistry* **73**, 1019-49(2004).

72. Crispin, M. et al. A human embryonic kidney 293T cell line mutated at the Golgi alpha-mannosidase II locus. *The Journal of biological chemistry* **284**, 21684-95(2009).

73. Gerken, T. a et al. Emerging paradigms for the initiation of mucin type protein O-glycosylation by the polypeptide GalNAc transferase (ppGalNAc T) family of glycosyltransferases. *The Journal of biological chemistry* **286**, 14493-14507(2011).

74. Gill, D.J. et al. Regulation of O-glycosylation through Golgi-to-ER relocation of initiation enzymes. *The Journal of cell biology* **189**, 843-58(2010).

75. Jínek, M. et al. The superhelical TPR-repeat domain of O-linked GlcNAc transferase exhibits structural similarities to importin alpha. *Nature structural & molecular biology* **11**, 1001-7(2004).

76. Iyer, S.P.N. & Hart, G.W. Roles of the tetratricopeptide repeat domain in O-GlcNAc transferase targeting and protein substrate specificity. *The Journal of biological chemistry* **278**, 24608-16(2003).

77. Ohn, T. et al. A functional RNAi screen links O-GlcNAc modification of ribosomal proteins to stress granule and processing body assembly. *Nature cell biology* **10**, 1224-31(2008).

78. Bionda, C. et al. Subcellular compartmentalization of ceramide metabolism: MAM (mitochondria-associated membrane) and/or mitochondria? *The Biochemical journal* **382**, 527-33(2004).

79. Ardail, D., Popa, I. & Bodennec, J. THE MITOCHONDRIA-ASSOCIATED-ER SUBCOMPARTMENT (MAM FRACTION) OF RAT LIVER CONTAINS HIGHLY ACTIVE SPHINGOLIPID-SPECIFIC GLYCOSYLTRANSFERASES. *Biochemical Journal* (2003).

80. Parsons, J.T., Horwitz, A.R. & Schwartz, M. a Cell adhesion: integrating cytoskeletal dynamics and cellular tension. *Nature reviews. Molecular cell biology* **11**, 633-43(2010).

Vita

Kanoelani Takaishi Pilobello was born in Hawaii to Kinue and Edwin around the same time that Mount St. Helens erupted. Her long format birth certificate is currently unavailable. Her parents received amnesty during the Regan administration allowing her to continue her education in the United States. This began at Edison Gifted Elementary School in Chicago, IL, which was the result of her mother refusing to send her to the local elementary school after witnessing questionable punishment of a child. She survived the inner city and more impressively secondary education in suburban Washington State. With the help of a number of grants from the college, she graduated from Bates College (Lewiston, ME) in 2001 with a degree in chemistry. She spent a year and a half at the Naval Research Laboratory in Washington, DC. In 2003, she came to The University of Texas at Austin through the Chemistry and Biochemistry program. In 2006, she entered the Institute of Cellular and Molecular Biology.

Permanent address (or email): ktpster@gmail.com

This dissertation was typed by the author.